目　录

第1章 数学预备知识

1.0 概述

本章从回顾微积分的一些重要的主题开始，这些主题在后继的章节中将会用到. 我们鼓励读者大胆地略过已熟悉的内容. 事实上，有些人也许希望从第2章开始.

1.1 基本概念和泰勒定理

首先回顾微积分的一些基本概念. 有人可能要问：我们只是对科学计算和数值算法感兴趣，为什么还要讨论这样的主题？熟知基本数学概念是理解大多数数值算法由来的基础. 但是，高深的数学概念并非必要，各种形式的泰勒定理既是许多数值计算过程的基础，又是研究科学计算的极佳切入点.

1.1.1 极限、连续性和导数

若 f 是一元实变函数，则函数 f 在 c 处的**极限**(如果存在)定义如下：等式

$$\lim_{x\to c} f(x) = L$$

表示对于每个正数 ϵ，存在对应的一个正数 δ，使得当 x 与 c 的距离小于 δ 时，$f(x)$ 与 L 的距离就小于 ϵ，即

$$\text{当 } 0 < |x-c| < \delta \text{ 时，有 } |f(x)-L| < \epsilon$$

若具有这样的性质的 L 不存在，则 f 在 c 处的极限就不存在.

例如，考虑函数

$$f(x) = x^2$$

由图1-1中 $f(x)=x^2$ 的图形显示，当 x 接近2时，$f(x)$ 接近4，因而等式

$$\lim_{x\to 2} x^2 = 4$$

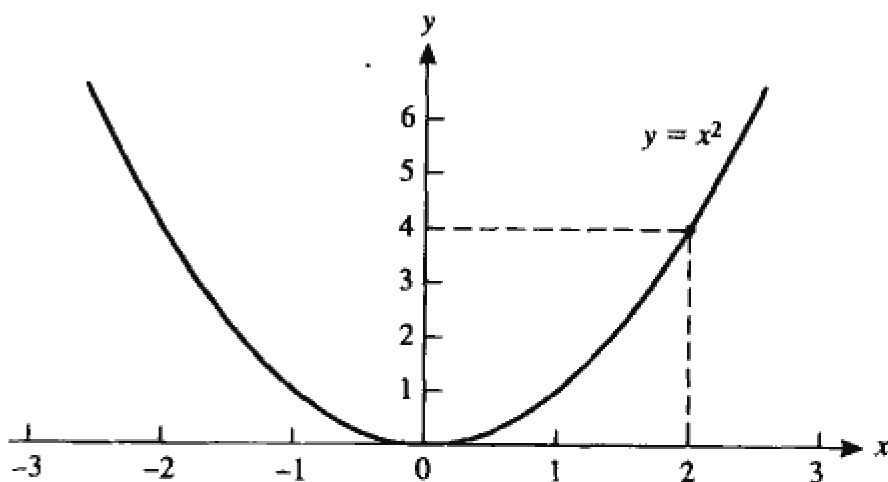


图1-1 $y=f(x)=x^2$

成立．现在我们通过证明：对 $\epsilon>0$，存在一个 $\delta>0$，使得当 $0<|x-2|<\delta$ 时，有 $|x^2-4|<\epsilon$，来领会为什么是这样的．设 $\epsilon>0$，$\delta=-2+\sqrt{4+\epsilon}>0$，于是 $\delta(\delta+4)=\epsilon$．若 $0<|x-2|<\delta$，则 $|x+2|=|x-2+4|\leqslant|x-2|+4<\delta+4$．从而，我们有 $|x^2-4|=|x-2||x+2|<(\delta+4)\delta=\epsilon$．注意，在某种意义上，为了发现怎样的 δ 值恰好表示 ϵ，我们逆向操作．显然，对于其他的 δ 值，如 $\delta=\epsilon/(5+\epsilon)$，也可以如此．（见习题 1.1.1.）

又例如，我们考虑

$$g(x)=\frac{|x|}{x}=\begin{cases}1 & 若\ x>0\\ -1 & 若\ x<0\end{cases}$$

图 1-2 是函数 $g(x)=|x|/x$ 的图形，从图中不难发现为什么 $g(x)$ 在 0 处没有定义．我们注意到等式

$$\lim_{x\to 0}\frac{|x|}{x}=L$$

不是对任意数 L 都成立的．实际上，设 $\epsilon=1$，假设当 $0<|x|<\delta$ 时，有 $||x|/x-L|<1$．若 $x=\delta/2$，则 $0<|x|<\delta$ 并且 $|x|/x=1$．但是若 $x=-\delta/2$，则 $0<|x|<\delta$ 并且 $|x|/x=-1$．在这两种情况下，均应该有 $||x|/x-L|<1$．然而没有既满足 $|1-L|<1$ 又满足 $|-1-L|<1$ 的这种数 L．因为这需要 L 同时满足 $0<L<2$ 和 $-2<L<0$．显然，这种情况是不可能的！所以我们说这个极限不存在．

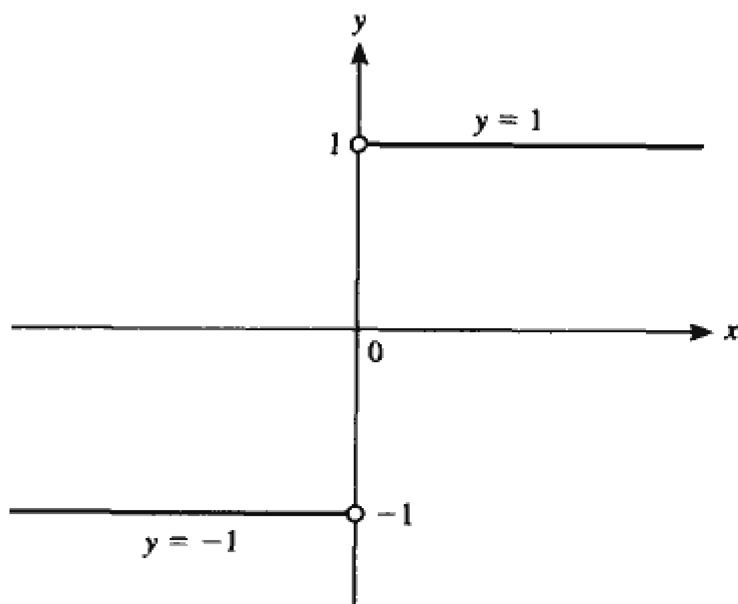


图 1-2　$y=g(x)=|x|/x$

若 f 仅定义在实轴的一个给定子集 X 上，则极限定义被修正为当 $x\in X$ 并且 $0<|x-c|<\delta$ 时，有 $|f(x)-L|<\epsilon$．

若

4

$$\lim_{x\to c}f(x)=f(c)$$

则称函数 f 在点 c 处**连续**．于是，函数 $f(x)=x^2$ 在点 2 处连续，然而函数 $|x|/x$ 在点 0 处不管是否有定义，它在点 0 处不连续．由前面所做的说明可得到这些论断．

下面给出一个直观上显而易见的定理．

定理 1(连续函数的介值定理) 定义在区间$[a, b]$上的连续函数 $f(x)$能取到 $f(a)$和 $f(b)$之间的所有值.

函数 f 在 c 处的**导数**(如果存在)由下式定义：

$$f'(c)=\lim_{x\to c}\frac{f(x)-f(c)}{x-c}$$

因为对于某个函数和特定的 c，这个极限可能不存在，所以对于这样的函数，可能导数不存在. 若函数 f 的$f'(c)$存在，则我们说 f 在 c 处**可微**. 若 f 在 c 处可微，则 f 必在 c 处连续. 现在来看看这是为什么. 考虑

$$\begin{aligned}\lim_{x\to c}[f(x)-f(c)] &= \lim_{x\to c}\frac{f(x)-f(c)}{x-c}(x-c)\\ &= f'(c)\cdot\lim_{x\to c}(x-c)=f'(c)\cdot 0=0\end{aligned}$$

显然，若 $f(x)$在 c 处可微，则$f'(x)$存在，并且$\lim_{x\to c}f(x)=f(c)$.

但是逆命题不成立！例如，若

$$f(x)=|x|$$

则$f'(0)$不存在. 见图 1-3 中 $f(x)=|x|$ 的图形. 在点 x 处的导数是曲线在 $f(x)$处的切线. 但是在“V”型曲线的底部($x=0$ 处)，不存在唯一的切线，因而在 $x=0$ 处没有导数.

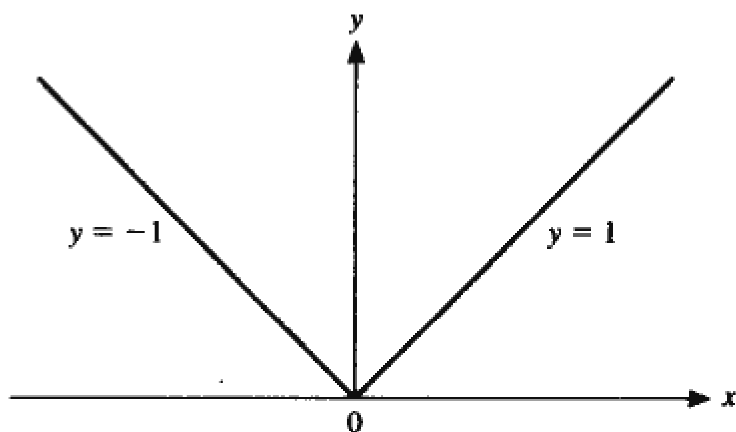


图 1-3 $y=f(x)=|x|$

在整个实轴$\mathbb{R}$ 上所有连续函数的集合记作$C(\mathbb{R})$. 所有导数处处连续的函数的集合记作$C^1(\mathbb{R})$. 若$f\in C^1(\mathbb{R})$，则f'在$\mathbb{R}$中所有点上连续而且 f 处处可微. 因为函数在一个点处的可微性蕴涵其在该点处的连续性，所以 $C^1(\mathbb{R})\subset C(\mathbb{R})$. 由于存在许多连续但不可导的函数，比如像$f(x)=|x|$ 这样的函数，所以集合 $C^1(\mathbb{R})$是 $C(\mathbb{R})$的一个真子集.

5

我们用 $C^2(\mathbb{R})$表示所有二阶导数处处连续的函数集合. 同理，

$$C^2(\mathbb{R})\subset C^1(\mathbb{R})\subset C(\mathbb{R})$$

还有，因为存在一阶可微而二阶不可微的函数，例如 $f(x)=x^2\sin(1/x)$(见习题 1.1.3)，所以这些包含是真包含.

类似地，我们定义 $C^n(\mathbb{R})$(n 是任意自然数)是所有 n 阶导数连续的函数集合. 最后，$C^\infty(\mathbb{R})$表示所有各阶导数都连续的函数集合. 我们有

$$C^\infty(\mathbb{R})\subset\cdots\subset C^2(\mathbb{R})\subset C^1(\mathbb{R})\subset C(\mathbb{R})$$

一个众所周知的函数 $f(x)=e^x$ 属于集合 $C^\infty(\mathbb{R})$.

同样，我们定义 $C^n[a, b]$是在闭区间$[a, b]$上所有 n 阶导数存在并且连续的函数集合.

1.1.2 泰勒定理

在 $C^n[a, b]$中，有关函数的一个重要定理是泰勒定理，它在数值分析的研究或科学计算中数值算法的研究中到处出现.

定理 2(带拉格朗日余项的泰勒定理)　若 $f\in C^n[a, b]$，且 $f^{(n+1)}$在开区间(a, b)上存在，则对于闭区间$[a, b]$中任意点 c 和 x，

$$f(x)=\sum_{k=0}^{n}\frac{1}{k!}f^{(k)}(c)(x-c)^k+E_n(x) \tag{1}$$

其中误差项是

$$E_n(x)=\frac{1}{(n+1)!}f^{(n+1)}(\xi)(x-c)^{n+1}$$

6 其中点 ξ 位于 c 和 x 之间.

这里"ξ 位于 c 和 x 之间"意味着依赖于所涉及的 c 和 x 的特定值，不是 $c<\xi<x$ 就是 $x<\xi<c$.

当 $c=0$ 时，出现一个重要的特殊情况. (1)式变成 $f(x)$的**麦克劳林级数**:

$$f(x)=\sum_{k=0}^{n}\frac{1}{k!}f^{(k)}(0)x^k+E_n(x) \tag{2}$$

其中

$$E_n(x)=\frac{1}{(n+1)!}f^{(n+1)}(\xi)x^{n+1}$$

对于一些重要函数，我们能得到它们的泰勒级数，如

$$\sin x=\sum_{k=0}^{\infty}(-1)^k\frac{x^{2k+1}}{(2k+1)!} \qquad (-\infty<x<\infty)$$

$$\cos x=\sum_{k=0}^{\infty}(-1)^k\frac{x^{2k}}{(2k)!} \qquad (-\infty<x<\infty)$$

$$\ln(1+x)=\sum_{k=1}^{\infty}\frac{(-1)^{k-1}}{k}x^k \qquad (-1<|x|<\infty)$$

$$\frac{1}{1+x}=\sum_{k=0}^{\infty}(-1)^kx^k \qquad (-1<x<1)$$

这些例子中出现的级数是**幂级数**. (见 6.7 节.)

例 1　利用泰勒定理，确定函数 $f(x)=\ln x$ 的泰勒级数，取 $a=1$，$b=2$，$c=1$.

解　公式需要 $f(x)=\ln x$ 的各阶导数，它们是$f'(x)=x^{-1}$，$f''(x)=-x^{-2}$，$f'''(x)=2x^{-3}$，$f^{(4)}(x)=-6x^{-4}$，等等. 接下来，我们得到通项[⊖]

$$f^{(k)}(x)=(-1)^{k-1}(k-1)!x^{-k} \qquad (k\geqslant 1)$$

显然，在 $x=1$ 处，我们有

⊖ 全书中，当涉及整数值的不等式 $1\leqslant i\leqslant m$ 时，意思是 $i=1, 2, \cdots, m$，类似地，$n\geqslant N$ 意思是 $n=N, N+1, \cdots$.

$$f^{(k)}(1) = (-1)^{k-1}(k-1)! \qquad (k \geqslant 1)$$

[7]

当然，$f^{(0)}(1)=f(1)=\ln 1=0$. 把所有导数值代入泰勒公式(1)，得到

$$\ln x = \sum_{k=1}^{n}(-1)^{k-1}\frac{1}{k}(x-1)^k + E_n(x) \qquad (1 \leqslant x \leqslant 2)$$

其中

$$E_n(x) = (-1)^n \frac{1}{n+1}\xi^{-(n+1)}(x-1)^{n+1} \qquad (1 < \xi < x)$$

■

在 $\ln x$ 的等式中，等号右边的和式 $\sum_{k=1}^{n}$ 产生如下所示的一个关于 x 的多项式函数. 它被认为是一个逼近较复杂函数 $\ln x$ 的简单函数. 右边最后一项 $E_n(x)$ 被认为是一个误差项. 它告诉我们多项式近似与 $\ln x$ 的差别. 注意这项不是一个多项式，因为 ξ 是以非多项式形式而依赖于 x 的.

可用泰勒公式计算函数在特定点的近似值. 例如，写出 $\ln x$ 展开式，我们有

$$\ln x = (x-1) - \frac{1}{2}(x-1)^2 + \frac{1}{3}(x-1)^3 - \cdots + (-1)^{n-1}\frac{1}{n}(x-1)^n + E_n(x)$$

其中

$$|E_n(x)| = \frac{1}{n+1}\xi^{-(n+1)}(x-1)^{n+1} < \frac{1}{n+1}(x-1)^{n+1}$$

在这个估计式中，我们只需指出 $1<\xi$ 以及 $\xi^{-(n+1)}<1$.

例 2 计算 $\ln 2$ 使其具有 $1/10^8$ 的精度，需要用级数的前多少项？

解 设 $x=2$，则我们有

$$\begin{aligned}\ln 2 &= 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots + (-1)^{n-1}\frac{1}{n} + E_n(2) \\ &= \sum_{k=1}^{n}(-1)^{k-1}\frac{1}{k} + E_n(2)\end{aligned}$$

其中 $|E_n(2)|<1/(n+1)$. $E_n(2)$项是数值误差. 因此为了计算 $\ln 2$ 使其具有期望的精度，需要选择 n 使得 $E_n(2)\leqslant 10^{-8}$. 这意味着 $1/(n+1)\leqslant 10^{-8}$ 或 $n+1\geqslant 10^8$. 因此在该多项式中至少需要 1 亿项来计算 $\ln 2$ 才能达到期望的精度！我们断定用泰勒级数计算 $\ln 2$ 不是一个实用的方法，需要采用另外的措施. 事实上，一个类似的计算表明仅用 22 项计算 $\ln 1.5$ 就能达到同样的精度.（见习题 1.1.5.）

■ [8]

$n=0$ 时的泰勒定理常常用于数学中的证明，称之为**中值定理**.

定理 3（中值定理） *若 f 属于 $C[a, b]$ 并且 f' 在开区间 (a, b) 上存在，则对于闭区间 $[a, b]$ 中的 x 和 c，*

$$f(x) = f(c) + f'(\xi)(x-c)$$

其中 ξ 位于 c 和 x 之间.

取 $x=b$ 和 $c=a$，并且重新整理上式，可得到重要的等式

$$f(b) - f(a) = f'(\xi)(b-a), \quad 其中 \quad a < \xi < b$$

由中值定理可以得到 $f'(x)$ 的一个近似式：

$$f'(x) \approx \frac{f(x+h)-f(x-h)}{2h}$$

这将在7.1节中讨论.

中值定理的一个特殊情况是**罗尔定理**.

定理4(罗尔定理) 若f在$[a,b]$上连续，f'在(a,b)上存在，并且$f(a)=f(b)$，则对于开区间(a,b)中的某个ξ，有$f'(\xi)=0$.

这是上面等式的直接结果.（实际上按常规的讨论，首先证明罗尔定理，然后从它导出泰勒定理.）在罗尔定理和中值定理中，区间$[a,b]$中可能存在不止一点ξ满足给定的等式.

在7.6节，我们将需要下面给出的另一种形式的泰勒定理. 它有几个良好的特性——证明简单明了并且带拉格朗日余项的泰勒定理可直接从带积分余项的泰勒定理得到.

定理5(带积分余项的泰勒定理) 若$f \in C^{n+1}[a,b]$，则对于闭区间$[a,b]$中的任意点x和c，

$$f(x) = \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(c)(x-c)^k + R_n(x) \tag{3}$$

其中

9

$$R_n(x) = \frac{1}{n!}\int_c^x f^{(n+1)}(t)(x-t)^n \mathrm{d}t$$

证明 回忆分部积分公式

$$\int u\mathrm{d}v = uv - \int v\mathrm{d}u$$

并且取

$$u = \frac{(x-t)^n}{n!} \quad \mathrm{d}v = f^{(n+1)}(t)\mathrm{d}t$$

应用到积分R_n上，得

$$\begin{aligned} R_n &= \frac{1}{n!}\left[f^{(n)}(t)(x-t)^n \Big|_{t=c}^{t=x} + n\int_c^x f^{(n)}(t)(x-t)^{n-1}\mathrm{d}t \right] \\ &= -\frac{1}{n!} f^{(n)}(c)(x-c)^n + R_{n-1} \end{aligned}$$

如果重复这个分部积分过程，我们最终得到

$$R_n = -\sum_{k=1}^{n} \frac{1}{k!} f^{(k)}(c)(x-c)^k + R_0$$

因为

$$R_0 = \int_c^x f'(t)\mathrm{d}t = f(x) - f(c)$$

所以我们有

$$f(x) = f(c) + \sum_{k=1}^{n} \frac{1}{k!} f^{(k)}(c)(x-c)^k + R_n$$

证毕. ■

1.1.3 泰勒公式的其他形式

在带拉格朗日余项的泰勒定理中，如果用 $x+h$ 和 x 分别替换 x 和 c，还能得到泰勒公式中级数与误差项的另一种形式. 这个结论如下.

定理 6(泰勒定理的另一种形式) 若 $f\in C^{n+1}[a, b]$，则对于闭区间$[a, b]$中的任意点 x 和 $x+h$，

$$f(x+h)=\sum_{k=0}^{n}\frac{h^k}{k!}f^{(k)}(x)+E_n(h) \tag{4}$$

其中

$$E_n(h)=\frac{h^{n+1}}{(n+1)!}f^{(n+1)}(\xi)$$

其中点 ξ 位于 x 和 $x+h$ 之间.

10

详细地，(4)式是

$$f(x+h)=f(x)+hf'(x)+\frac{h^2}{2}f''(x)+\frac{h^3}{3!}f'''(x)+\cdots+\frac{h^n}{n!}f^{(n)}(x)+E_n(h)$$

对于许多应用来说，这是一个很重要的形式.

例 3 确定 A^{x+h} 的泰勒公式并且近似计算 $10^{1.0001}$.

解 设 $f(x)=A^x$，有 $f^{(n)}(x)=A^x(\ln A)^n$. 利用(4)式，我们有

$$A^{x+h}=A^x\left(1+\sum_{k=1}^{n}\frac{h^k}{k!}(\ln A)^k\right)+E_n(h)$$

下面取 $A=10$，$x=1$，$h=10^{-4}$. 从而得到

$$\begin{aligned}10^{1.0001}&=10(1+10^{-4}(\ln 10)+\frac{1}{2}10^{-8}(\ln 10)^2+\cdots)\\&\approx 10(1+2.30259\times 10^{-4}+2.65095\times 10^{-8})\\&\approx 10.0023000265095\end{aligned}$$

■

对向量值函数，也存在泰勒级数和泰勒公式. 若 f 是一个$\mathbb{R}^n$ 到$\mathbb{R}^m$ 的映射，则用 $f(x)$，$f'(x)$，$f''(x)$等表示 $f(x+h)$的公式是有效的. 当然，主要困难在于定义适当的导数. 这些内容已收录在许多教科书中，例如 Bartle[1976]，K. T. Smith[1971]和 Dieudonné[1960]. 下面我们将提及一些在后面章节中需要的某些特殊情况.

对于函数 f: $\mathbb{R}^2\to\mathbb{R}$，泰勒公式最简单的表达式是一个符号表达式：

定理 7(二元泰勒定理) 设 $f\in C^{n+1}([a, b]\times[c, d])$. 若$(x, y)$和$(x+h, y+k)$是矩形$[a, b]\times[c, d]\subseteq\mathbb{R}^2$ 内的点，则

$$f(x+h,y+k)=\sum_{i=0}^{n}\frac{1}{i!}\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y}\right)^i f(x,y)+E_n(h,k) \tag{5}$$

其中

$$E_n(h,k)=\frac{1}{(n+1)!}\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y}\right)^{n+1}f(x+\theta h,y+\theta k)$$

其中 θ 位于 0 和 1 之间.

11

本定理中难以理解的项是指：

$$\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y}\right)^0 f(x,y)=f(x,y)$$

$$\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y}\right)^1 f(x,y)=\left(h\frac{\partial f}{\partial x}+k\frac{\partial f}{\partial y}\right)(x,y)$$

$$\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y}\right)^2 f(x,y)=\left(h^2\frac{\partial^2 f}{\partial x^2}+2hk\frac{\partial^2 f}{\partial x\partial y}+k^2\frac{\partial^2 f}{\partial y^2}\right)(x,y)$$

等项．设 $f_x=\partial f/\partial x$，$f_y=\partial f/\partial y$，$f_{xx}=\partial^2 f/\partial x^2$，$f_{xy}=\partial^2 f/\partial x\partial y$，$f_{yy}=\partial^2 f/\partial y^2$，则我们可把(5)式的前几项写成

$$f(x+h,y+k)=f+(hf_x+kf_y)+\frac{1}{2}(h^2 f_{xx}+2hkf_{xy}+k^2 f_{yy})+\cdots$$

其中等式右边的函数 f 和其后的每个偏导数都取$(x,\ y)$处的值．

例 4 函数 $f(x,\ y)=\cos(xy)$的泰勒公式中前几项是什么？

解 对于给定的函数，我们求出

$$\frac{\partial f}{\partial x}=-y\sin(xy)\qquad \frac{\partial f}{\partial y}=-x\sin(xy)$$

$$\frac{\partial^2 f}{\partial x^2}=-y^2\cos(xy)\qquad \frac{\partial^2 f}{\partial x\partial y}=-xy\cos(xy)-\sin(xy)\qquad \frac{\partial^2 f}{\partial y^2}=-x^2\cos(xy)$$

于是，在泰勒公式(5)中，若取 $n=1$，则结果是

$$\cos[(x+h)(y+k)]=\cos(xy)-hy\sin(xy)-kx\sin(xy)+E_1(h,k)$$

余项 E_1 是 3 项之和，即

$$-\frac{1}{2}h^2(y+\theta k)^2\cos[(x+\theta h)(y+\theta k)]$$

$$-hk\{(x+\theta h)(y+\theta k)\cos[(x+\theta h)(y+\theta k)]+\sin[(x+\theta h)(y+\theta k)]\}$$

$$-\frac{1}{2}k^2(x+\theta h)^2\cos[(x+\theta h)(y+\theta k)]$$

■

习题 1.1

1. 证明当 $0<|x-2|<\varepsilon(5+\varepsilon)^{-1}$时有 $|x^2-4|<\varepsilon$ 并且用这些不等式证明$\lim\limits_{x\to 2}x^2=4$.
2. 函数 $f(x)=x\sin(1/x)$，其中 $f(0)=0$，证明 $f(x)$在 0 处连续但不可微.
3. [12] $f(x)=x^2\sin(1/x)$，其中 $f(0)=0$，证明 $f(x)$在 0 处一阶可微但二阶不可微.
4. 设 $f(x)=x^{-3}(x-\sin x)$，$x\neq 0$. 为了使 f 连续，$f(0)$应该怎样定义？要使 f 还是可微的，$f(0)$又该怎样定义？
5. a. 导出函数 $f(x)=\ln(x+1)$在 0 处的泰勒级数．用求和记号写出这个级数．当这个级数被截断时，给出余项的两个表达式.

   b. 确定级数最少需要多少项才能使计算 ln1.5 的误差小于 10^{-8}.

   c. 确定级数需要多少项才能使计算 ln1.6 的误差至多为 10^{-10}.
6. 判断下列函数是否连续？一阶可微或二阶可微？

$$f(x)=\begin{cases}x^3+x-1 & 若\ x\leqslant 0\\ x^3-x-1 & 若\ x>0\end{cases}$$

7. 对下列函数重复讨论上述问题：

$$f(x)=\begin{cases}x & 若 x\leqslant 1\\ x^2 & 若 x>1\end{cases}$$

8. 判断下列推理是否正确：

函数

$$f(x)=\begin{cases}x^3+x & 若 x\leqslant 0\\ x^3-x & 若 x\geqslant 0\end{cases}$$

具有性质

$$\lim_{x\to 0^+} f''(x)=\lim_{x\to 0^+} 6x=0$$
$$\lim_{x\to 0^-} f''(x)=\lim_{x\to 0^-} 6x=0$$

所以，f''是连续的.

9. 证明：若 f 在 x 处可微，则

$$\lim_{h\to 0}\frac{f(x+h)-f(x-h)}{2h}=f'(x)$$

举例说明，存在某些在 x 处不可微的函数，但是上面的极限存在.（见 Eggermont[1988]或下一题.）

10. 证明或否定论断：若 f 在 x 处可微，则对于 $\alpha\neq 1$，有

$$f'(x)=\lim_{h\to 0}\frac{f(x+h)-f(x+\alpha h)}{h-\alpha h}$$

11. 用 ϵ-δ 证法证明 $\lim\limits_{x\to 1}(4x+2)=6$.
12. 用 ϵ-δ 证法证明 $\lim\limits_{x\to 2}(1/x)=\frac{1}{2}$.
13. 对于函数 $f(x)=3-2x+x^2$ 和区间$[a, b]=[1, 3]$，求中值定理中出现的 ξ.
14. （续）对于函数 $f(x)=x^6+x^4-1$ 和区间$[a, b]=[0, 1]$，重复讨论上述问题.
15. 求 $f(x)=\cosh x$ 在点 $c=0$ 处的泰勒级数.
16. 如果 $\ln x$ 的级数在含$(x-1)^{1000}$的项后被截断，然后用它来计算 $\ln 2$，那么能给出怎样的误差界？ 13
17. 求 $f(x)=e^x$ 在点 $c=3$ 处的泰勒级数. 然后化简这个级数，并且说明如何从 f 在点 $c=0$ 处的级数直接得到此级数.
18. 设 k 是一个正整数并且 $0<\alpha<1$. 函数 $x^{k+\alpha}$属于哪个类 $C^n(\mathbb{R})$？
19. 证明：若 $f\in C^n(\mathbb{R})$，则$f'\in C^{n-1}(\mathbb{R})$且$\int_a^x f(t)\mathrm{d}t\in C^{n+1}(\mathbb{R})$.
20. 直接证明罗尔定理（不作为中值定理的特殊情况）.
21. 证明：若 $f\in C^n(\mathbb{R})$并且 $f(x_0)=f(x_1)=\cdots=f(x_n)=0$，$x_0<x_1<\cdots<x_n$，则对于某个 $\xi\in(x_0, x_n)$，$f^{(n)}(\xi)=0$. 提示：使用 n 次罗尔定理.
22. 证明函数 $f(x)=x^2$ 处处连续.
23. 对于小值 x，常常使用近似 $\sin x\approx x$. 借助于泰勒定理估计使用该公式产生的误差. 求 x 值的范围，使得该近似给出精确到 6 位小数的结果.
24. 对于小值 x，近似 $\cos x\approx 1-\frac{1}{2}x^2$ 的有效性怎样？求 x 值的范围，使得该近似给出精确到 3 位小数的结果.
25. 用 $n=2$ 的泰勒定理证明对所有非零实数 x，有不等式 $1+x<e^x$.
26. 对 $\ln(1+x)$在 1 处导出带余项的泰勒级数. 导出一个关于泰勒级数项数的不等式，为使得计算 $\ln 4$ 时误差

小于 2^{-m}，给出所需的项数.

27. x^2+x-2 在点 3 处的泰勒展开式中的第 3 项是什么?

28. 用 e^x 的级数计算 e^2，需要多少项才能精确到 4 位小数(舍入)?

29. 在 e 处把 $f(x)=\ln x$ 展开成泰勒级数，写出用和式记号的结果并给出余项. 假设 $|x-e|<1$ 并且精度为 $\frac{1}{2}\times 10^{-1}$. 试问为了达到这个精度，最少需要该级数多少项?

30. 求 x^x 在 1 处泰勒级数的前 2 项及余项 E_1.

31. 求 $f(x)=e^{(\cos x)}$ 在点 π 处展开的 2 次泰勒多项式.

32. 首先把函数$\sqrt{x}$以$(x-1)$的幂级数形式展开，然后用此级数把$\sqrt{0.999\,999\,999\,5}$近似到 10 位小数.

33. 假定 $|x|<\frac{1}{2}$. 利用泰勒定理求下式的最佳上界.

    a. $|\cos x-(1-x^2/2)|$

    b. $|\sin x-x(1-x^2/6)|$

34. 求一个被称为 x^3-2x 在 2 处**线性化**的函数.

35. 需要级数

$$e=\sum_{k=0}^{\infty}\frac{1}{k!}$$

的多少项才能使得 e 在第 20 个小数位上具有至多 6/10 个单位的误差?

36. 求 $x^{1/5}$ 在点 $x=32$ 处的泰勒展开的前 2 项. 用此级数的这 2 项近似 31.999 999 的 5 次方根. 试问你的答案有怎样的精度?

[14] 37. 求函数 $f(x)=e^{2x}\sin x$ 在点 $\pi/2$ 处展开的 2 次泰勒多项式.

38. 当泰勒定理被应用到函数 $f(x)=\cos x$，$n=2$ 且 $c=\pi/2$ 时，求拉格朗日形式的余项. 若余项的绝对值不超过$\frac{1}{2}\times 10^{-4}$，$|x-\pi/2|$ 必须取多小?

39. 在说明泰勒公式的例子中给出了$(-1)^n(n+1)^{-1}\xi^{-n-1}(x-1)^{n+1}$形式的误差项. 将这个误差项与由积分形式的余项引起的误差项进行比较.

40. 用带积分余项的泰勒定理和积分中值定理推导带拉格朗日余项的泰勒定理.

1.2　收敛阶及相关基本概念

在数值计算中，尤其是在高性能计算机上，一个问题的解答常常不会立即产生. 更确切地说，通常产生一列呈现精度逐渐提高的近似解. 序列收敛性是在后面，例如在第 3 章中继续讨论的一个重要课题，这里我们仅仅给一些介绍性的概念.

1.2.1　收敛序列

考虑一种理想的情况：一个问题只有唯一的实数解. 例如，它可能是复杂方程的零点或是难以处理的定积分的数值. 在这种情况下，计算机程序会产生一列逼近正确解的实数序列：x_1，x_2，x_3，$\cdots$.

对每个正数 ε 若存在一个实数 r，使得当 $n>r$ 时(这里 n 是一个整数)，有 $|x_n-L|<\varepsilon$，记作

$$\lim_{n\to\infty}x_n=L$$

例如，

$$\lim_{n\to\infty}\frac{n+1}{n}=1$$

这是因为当 $n>\varepsilon^{-1}$时，有

$$\left|\frac{n+1}{n}-1\right|<\varepsilon$$

又例如，回忆用来定义重要无理数 e 的等式

[15]

$$\mathrm{e}=\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n$$

若我们计算序列 $x_n=(1+1/n)^n$，则这些元素是

$$x_1 = 2.000\,000$$
$$x_{10} = 2.593\,742$$
$$x_{30} = 2.674\,319$$
$$x_{50} = 2.691\,588$$
$$x_{1\,000} = 2.716\,924$$

因为这个序列的极限是 e=2.718 281 8…，并且其第 1 000 项的误差还是 0.001 358，因此它是一个收敛速度相当慢的序列的实例. 用双精度计算，我们发现数值征兆

$$\frac{|x_{n+1}-\mathrm{e}|}{|x_n-\mathrm{e}|}\to 1$$

这个性质比**线性收敛**(稍后精确定义)差些.

一个以略微较快的速度收敛于$\sqrt{2}$的序列的实例是

$$x_{n+1}=x_n-(x_n^2-2)\frac{x_n-x_{n-1}}{x_n^2-x_{n-1}^2}$$

选择两个初始值，我们有

$$x_1 = 2$$
$$x_2 = 1.5$$
$$x_3 = 1.428\,571$$
$$x_4 = 1.414\,634$$
$$x_5 = 1.414\,216$$
$$x_6 = 1.414\,214$$

很快地收敛于$\sqrt{2}=1.414\,213\,562\cdots$. 利用双精度计算我们发现对应于**超线性收敛**的数值征兆

$$\frac{|x_{n+1}-\sqrt{2}|}{|x_n-\sqrt{2}|^{1.62}}\leqslant 0.77$$

作为快速收敛序列的实例，考虑由下列递归关系定义的序列：

[16]

$$\begin{cases}x_1=2\\ x_{n+1}=\dfrac{1}{2}x_n+\dfrac{1}{x_n}\quad(n\geqslant 1)\end{cases}$$

这个序列的元素是

$$x_1 = 2.000\,000$$
$$x_2 = 1.500\,000$$
$$x_3 = 1.416\,667$$
$$x_4 = 1.414\,216$$

极限是$\sqrt{2}=1.414\,213\,562\cdots$，并且该序列以极快的速度收敛到其极限．用双精度计算，我们发现征兆

$$\frac{|x_{n+1}-\sqrt{2}|}{|x_n-\sqrt{2}|^2} \leqslant 0.36$$

这样的条件对应于我们不久将看到的**二阶收敛**．

1.2.2 收敛阶

用一些特殊的术语来描述序列收敛的速度．设$[x_n]$是一个趋于极限x^*的实数序列．若存在一个常数$c<1$和一个整数N使得

$$|x_{n+1}-x^*| \leqslant c\,|x_n-x^*| \quad (n \geqslant N)$$

则我们说序列的收敛速度至少是**线性**的．若存在一个趋向于0的序列ε_n和一个整数N使得

$$|x_{n+1}-x^*| \leqslant \varepsilon_n\,|x_n-x^*| \quad (n \geqslant N)$$

则我们说序列的收敛速度至少是**超线性**的．若存在一个常数C(不必小于1)和一个整数N使得

$$|x_{n+1}-x^*| \leqslant C\,|x_n-x^*|^2 \quad (n \geqslant N)$$

则我们说序列的收敛速度至少是**二阶**的．一般地，如存在正常数C和α以及一个整数N使得

$$|x_{n+1}-x^*| \leqslant C\,|x_n-x^*|^\alpha \quad (n \geqslant N)$$

则我们说序列的收敛速度至少是**α阶**的．

1.2.3 大$\mathcal{O}$和小o记号

现在考虑几种比较两个序列或两个函数的标准方法．我们从序列开始．

设$[x_n]$和$[\alpha_n]$是两个不同序列．若存在常数C和n_0，使得当$n\geqslant n_0$时，有$|x_n|\leqslant C|\alpha_n|$，则记

[17]

$$x_n = \mathcal{O}(\alpha_n)$$

这里我们说x_n是α_n的“大$\mathcal{O}$”．

等式

$$x_n = o(\alpha_n)$$

直观地表示$\lim\limits_{n\to\infty}(x_n/\alpha_n)=0$的意思．这里我们说$x_n$是$\alpha_n$的“小$o$”．为避免被零除，严谨的定义应当是对于某个$\varepsilon_n\geqslant 0$，我们有$\varepsilon_n\to 0$并且$|x_n|\leqslant\varepsilon_n|\alpha_n|$．

这两个概念给出了一种比较两个序列的粗糙方法．当两个序列都收敛于0时，我们经常用到它们．若$x_n\to 0$，$\alpha_n\to 0$，并且$x_n=\mathcal{O}(\alpha_n)$，则$x_n$收敛于0的速度至少与$\alpha_n$一样快．若$x_n=o(\alpha_n)$，则$x_n$收敛于0的速度比$\alpha_n$快．

下面给出一些实例：

$$\frac{n+1}{n^2}=\mathcal{O}\left(\frac{1}{n}\right) \tag{1}$$

$$\frac{1}{n\ln n}=o\left(\frac{1}{n}\right) \tag{2}$$

$$\frac{1}{n}=o\left(\frac{1}{\ln n}\right) \tag{3}$$

$$\frac{5}{n}+\mathrm{e}^{-n}=\mathcal{O}\left(\frac{1}{n}\right) \tag{4}$$

$$\mathrm{e}^{-n}=o\left(\frac{1}{n^2}\right) \tag{5}$$

对于 1.1 节的例题，我们能写成

$$\ln 2-\sum_{k=1}^{n-1}(-1)^{k-1}\frac{1}{k}=\mathcal{O}\left(\frac{1}{n}\right) \tag{6}$$

这是一个非常**慢收敛**的实例. 另一方面，

$$\mathrm{e}^x-\sum_{k=0}^{n-1}\frac{1}{k!}x^k=\mathcal{O}\left(\frac{1}{n!}\right) \quad (|x|\leqslant 1) \tag{7}$$

这是一个非常**快收敛**的实例.

刚才介绍的符号除了用于序列外，也可用于函数. 例如，我们能写

$$\sin x=x-\frac{x^3}{6}+\mathcal{O}(x^5) \quad (x\to 0) \tag{8}$$

18

这个式子意指存在一个 0 的邻域和一个常数 C，使得在该邻域上，有

$$\left|\sin x-x+\frac{x^3}{6}\right|\leqslant C|x^5|$$

这个论断的正确性可以用 $n=4$ 和 $f(x)=\sin x$ 的泰勒定理来验证.

等式

$$f(x)=\mathcal{O}(g(x)) \quad (x\to\infty)$$

意味着存在常数 r 和 C 使得当 $x\geqslant r$ 时，有 $|f(x)|\leqslant C|g(x)|$. 例如，因为当 $x\geqslant 1$ 时，有 $\sqrt{x^2+1}\leqslant 2x$，于是有

$$\sqrt{x^2+1}=\mathcal{O}(x) \quad (x\to\infty)$$

在使用符号 $f(x)=\mathcal{O}(g(x))$或 $f(x)=o(g(x))$时，重要的是要说明有关收敛的点是什么. 例如，当 $x\to\infty$时，$x^{-2}=o(x^{-1})$，但是在 0 处关系相反：当 $x\to 0$ 时，$x^{-1}=o(x^{-2})$.

一般地，当存在一个常数 C 和一个 x^* 的邻域，使得在该邻域内有 $|f(x)|\leqslant C|g(x)|$ 时，我们记

$$f(x)=\mathcal{O}(g(x)) \quad (x\to x^*)$$

类似地，

$$f(x)=o(g(x)) \quad (x\to x^*)$$

意指 $\lim\limits_{x\to x^*}[f(x)/g(x)]=0$.

1.2.4 积分中值定理

下面是在数值分析中常常使用的另一个中值定理.

定理1(积分中值定理) 设u和v是在区间$[a, b]$上连续的实函数，并且假设$v \geqslant 0$. 则在$[a, b]$内存在一点ξ，使得

$$\int_a^b u(x)v(x)\mathrm{d}x = u(\xi)\int_a^b v(x)\mathrm{d}x$$

证明 设α和β分别表示$u(x)$在$[a, b]$上的最小值和最大值. 因此有

$$\alpha \leqslant u(x) \leqslant \beta \quad (a \leqslant x \leqslant b)$$

因为$v(x) \geqslant 0$，所以我们有

19

$$\alpha v(x) \leqslant u(x)v(x) \leqslant \beta v(x)$$

现在对整个不等式积分，并且令$I = \int_a^b v(x)\mathrm{d}x$. 则结果是

$$\alpha I \leqslant \int_a^b u(x)v(x)\mathrm{d}x \leqslant \beta I$$

如果$I=0$，则$v(x) \equiv 0$，而且我们希望证明的结论是平凡的. 如果I不是0，则有

$$\alpha \leqslant I^{-1}\int_a^b u(x)v(x)\mathrm{d}x \leqslant \beta$$

根据连续函数的介值定理得，在$[a, b]$内存在一点ξ，使得

$$u(\xi) = I^{-1}\int_a^b u(x)v(x)\mathrm{d}x$$

■

1.2.5 嵌套乘法

多项式的嵌套乘法是本书中许多地方都需要的一个基本概念. 多项式能以嵌套形式重写使得当计算它的数值时，只需要略多于最低限度的乘法次数即可. 多项式

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0$$

能用包括和$\sum$与积Π的标准数学符号写成：

$$p(x) = \sum_{k=0}^{n} a_k x^k = \sum_{k=0}^{n}\left(a_k \prod_{j=1}^{k} x\right)$$

回忆若$n \leqslant m$，则

$$\sum_{k=n}^{m} s_k = s_n + s_{n+1} + \cdots + s_m, \prod_{k=n}^{m} y_k = y_n y_{n+1} \cdots y_m$$

因此

$$\sum_{k=n}^{m} r = (m-n+1)r, \prod_{k=n}^{m} x = x^{(m-n+1)}$$

为了方便，若$m<n$，则

$$\sum_{k=n}^{m} s_k = 0, \prod_{k=n}^{m} y_k = 1$$

要有效地求多项式的值，我们可以用**嵌套乘法**把项分组：

$$p(x) = a_0 + x(a_1 + x(a_2 + \cdots + x(a_{n-1} + x(a_n))\cdots))$$

上式对应于下面仅需要 n 次乘法和 n 次加法的简单算法： [20]

```
p←a_n
for k=n−1 to 0 step −1 do
    p←xp+a_k
end do
```

这个程序也称为**霍纳方法**或**综合除法**.

1.2.6 上界和下界

在数值分析中频繁出现的两个重要概念是**上确界(最小上界)**和**下确界(最大下界)**. 设 S 是一个非空的有界实数集. **有界性**是指：对于适当的实数 a 和 b，有

$$a \leqslant x \leqslant b, \text{对所有 } x \in S$$

在这种情况下，我们称 b 是 S 的一个上界，a 是 S 的一个下界. 当然，S 有许多上界：如果 $c \geqslant b$，那么 c 也是 S 的一个上界. 显然，集合 S 没有最大的上界，但有最小的上界.

公理 1(最小上界公理) 任何有上界的非空实数集都有一个最小上界.

公理中所描绘的性质是实数系较深刻的特征之一(例如，有理数系没有这个性质). 一个集合 S 的**最小上界**(如果存在一个)记作 $\text{lub}S$. 名词**上确界**与它同义，而且它被简写为 $\sup S$. 于是，我们可以给出以下定义：

定义 1(上确界的定义) S 的上确界是 $v(v=\sup S=\text{lub}S)$ 当且仅当

1. v 是 S 的一个上界.
2. S 的上界中没有比 v 小的实数.

例 1 若 $S=\{x : x^2<2\}$，试问 $\sup S$ 是多少?

解 因为 $S=\{x : -\sqrt{2}<x<\sqrt{2}\}$，所以 S 的最小上界是$\sqrt{2}$. ■

若 F 是一个函数，则符号$\sup\limits_{x\in A} f(x)$意指 $\sup\{f(x): x\in A\}$. 例如，我们可以证明

$$\sup_{0<x<\frac{\pi}{6}} \sin x = \frac{1}{2}$$

相应的**最大下界**或**下确界**概念，记为 $u=\text{glb}S$ 或 $u=\inf S$，其定义如下： [21]

定义 2(下确界的定义) S 的下确界是 $u(u=\inf S=\text{glb}S)$ 当且仅当

1. u 是 S 的一个下界.
2. S 的下界中没有比 u 大的实数.

当我们把公理应用到集合 $-S=\{-x: x\in S\}$ 上时，就得到一个基本结论：若一个非空实数集有一个下界，则它有一个最大的下界.

1.2.7 显函数与隐函数

函数通常凭借显式公式来定义，对每个自变量，可由这个显式公式计算出该函数值. 例如，我们可用公式

$$f(x) = \sqrt{7x^3 - 2x}$$

定义函数 f. 在扩大常用函数范围后，可定义像

$$f(x)=\ln(\arctan x)+\cos(e^x)$$

这种更复杂的函数. 另外还有许多定义函数的方法，比如，通过微分方程、积分、无穷级数等来定义函数. 因而可用下面带初始条件的微分方程

$$\begin{cases}y'=1+\sin y\\y(0)=0\end{cases}$$

适当地定义一个函数 $y=f(x)$. 另外一个例子称为**误差函数**，记为 $\mathrm{erf}(x)$，它是由下列积分定义的：

$$\mathrm{erf}(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt$$

在本节中，我们考虑**隐式**定义的函数. 这意味着给定一个有两个变量的函数 G，而且希望从方程 $G(x, y)=0$ 中重新获得 x 的函数 y. 有时，我们能解方程 $G(x, y)=0$ 得到 $y=f(x)$. 例如，从方程

$$y^2+3xy-7=0$$

中，可求得 y，因此重新获得两个显函数

$$y=\frac{1}{2}(-3x\pm\sqrt{9x^2+28})$$

同样，从方程

22

$$\sin(y+7)=x^3-2$$

中求 y，可重新获得一个显函数：

$$y=\sin^{-1}(x^3-2)-7$$

因为可选择反正弦函数的许多**分支**，所以在这里也存在多个函数.

一般而言，若 G 是给定的函数，并且在特定点 (x_0, y_0) 处 $G(x_0, y_0)=0$，则我们要求在附近的其他点也满足方程 $G(x, y)=0$. 这样，当 x 改变时，为使方程 $G(x, y)=0$ 成立，必然要求 y 作适当的变化. 所以，y 应该是 x 在 x_0 的某个邻域内的一个函数. 图 1-4 显示了这种典型的情况.

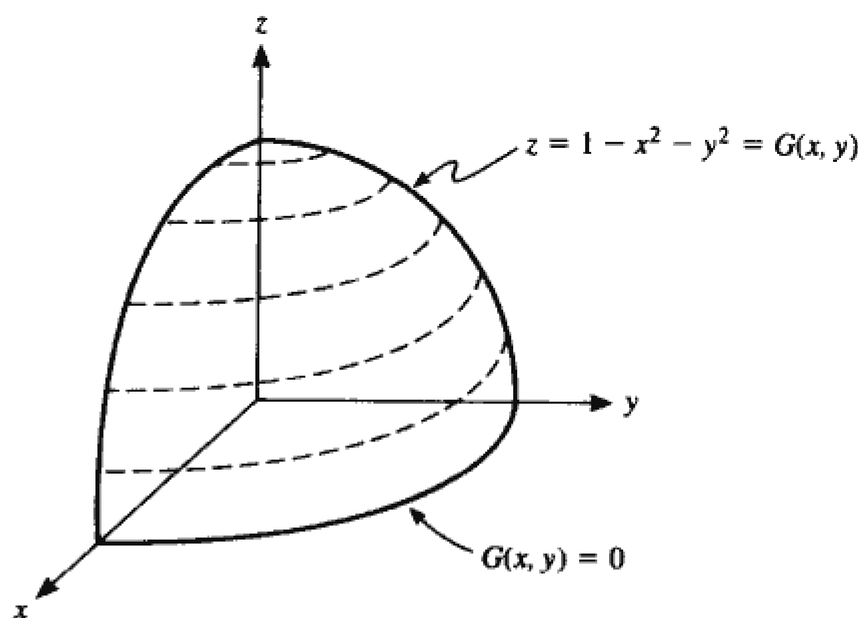


图 1-4 曲面 G 与 xy 面交于曲线 $G(x, y)=0$

定理 2(隐函数定理) 设 G 是一个有两个实变量的函数，它在(x_0, y_0)的一个邻域内有定义并且连续可微. 若 $G(x_0, y_0)=0$，并且在(x_0, y_0)处 $\partial G/\partial y\neq 0$，则存在一个正数 δ 和一个定义在 $|x-x_0|<\delta$ 上的连续可微函数 f，使得 $f(x_0)=y_0$ 和 $G(x, f(x))=0$.

例 2 试问方程

$$x^7+2y^8-y^3=0$$

是否可以在 $x=-1$ 的某个邻域内定义一个 x 的连续可微函数 y?

解 设$(x_0, y_0)=(-1, 1)$，并取

$$G(x,y)=x^7+2y^8-y^3$$

则 $G(x_0, y_0)=0$，并且

$$\frac{\partial G}{\partial y}=16y^7-3y^2$$

因此，$\partial G(x_0, y_0)/\partial y=13$. 根据隐函数定理，$y$ 是 x 在 $x_0=-1$ 附近连续可微的函数. ■ [23]

若 f 是一个由 $G(x, y)=0$ 隐式定义的函数，则对某个区间中的 x，$G(x, f(x))=0$. 若想计算$f'(x)$，则可以使用微积分中常见的技巧. 对方程 $G(x, y)=0$ 关于 x 求导，注意要把 y 理解为 x 的函数. 结果是

$$\frac{\partial G}{\partial x}+\frac{\partial G}{\partial y}\frac{dy}{dx}=0$$

由此得到

$$\frac{dy}{dx}=-\frac{\partial G}{\partial x}\Big/\frac{\partial G}{\partial y}$$

或

$$f'(x)=-\frac{\partial G}{\partial x}\Big/\frac{\partial G}{\partial y}$$

如此得到的导数很可能就是 x 和 y 的函数.

例 3 若 $y(x)$是由方程 $x^3-y^7+4x^2+y^4-24=0$ 隐式定义的，试问 dy/dx 在点(2, 1)的值是多少?

解 由关于 x 的全微分得到

$$3x^2-7y^6\frac{dy}{dx}+8x+4y^3\frac{dy}{dx}=0$$

把 $x=2$ 和 $y=1$ 代入上式，我们得到

$$12-7\frac{dy}{dx}+16+4\frac{dy}{dx}=0$$

因此可得 $dy/dx=28/3$. ■

一些涉及隐函数的数值问题在 3.2 节中讨论.

习题 1.2

1. 当 $x\to 0$ 时，确定方程

$$\arctan x=x+\mathcal{O}(x^k)$$

中 k 的最佳整数值.

2. 设序列 x_n 由 $x_{n+1}=F(x_n)$归纳地定义．若当 $n\to\infty$时，$x_n\to x$ 并且 $F'(x)=0$．证明

$$x_{n+2}-x_{n+1}=o(x_{n+1}-x_n)$$

提示：用中值定理并且假定 F 是连续可微函数．

[24] 3. 证明每个充分光滑的函数在长度为 h 的区间上能被 n 次多项式近似，并且当 $h\to 0$ 时误差是 $\mathcal{O}(h^{n+1})$．

4. a. 考虑级数

$$e^{\tan x}=1+x+\frac{x^2}{2!}+\frac{3x^3}{3!}+\frac{9x^4}{4!}+\cdots \quad (|x|\leqslant \pi/2)$$

保留级数中的3项，当 $x\to 0$ 时，用带有最佳整数值的 o 记号估计剩余的级数．

b. 对于级数

$$\ln\tan x=\ln x+\frac{x^2}{3}+\frac{7x^4}{90}+\frac{62x^6}{2\,835}+\cdots \quad (0<|x|<\pi/2)$$

用 $\mathcal{O}$ 记号重做上述问题．

5. 对于整数值 γ 和 δ，当 $x\to 0$ 时，用 $\mathcal{O}(x^\gamma)$或用 $o(x^\delta)$代替级数

$$\ln(1+x)=\sum_{k=1}^{n-1}(-1)^{k-1}\frac{x^k}{k}+\cdots$$

中的$+\cdots$．确定 γ 和 δ 的范围．

6. 对于数对(x_n,α_n)，当 $n\to\infty$时，$x_n=\mathcal{O}(\alpha_n)$是否成立？

a. $x_n=5n^2+9n^3+1$，$\alpha_n=n^2$

b. $x_n=5n^2+9n^3+1$，$\alpha_n=1$

c. $x_n=\sqrt{n+3}$，$\alpha_n=1$

d. $x_n=5n^2+9n^3+1$，$\alpha_n=n^3$

e. $x_n=\sqrt{n+3}$，$\alpha_n=1/n$

7. 判断下式是否正确(这里 $n\to\infty$)．

a. $(n+1)/n^2=o(1/n)$

b. $(n+1)/\sqrt{n}=o(1)$

c. $1/\ln n=\mathcal{O}(1/n)$

d. $1/(n\ln n)=o(1/n)$

e. $e^n/n^5=\mathcal{O}(1/n)$

8. 当 $h\to 0$ 时，表达式 e^h，$(1-h^4)^{-1}$，$\cos(h)$，$1+\sin(h^3)$有相同的极限．用下列具有最佳整数值 α 和 β 的形式分别表示这些表达式．

$$f(h)=c+\mathcal{O}(h^\alpha)=c+o(h^\beta)$$

9. (续)当 $h\to 0$ 时，试问下列表达式的极限和收敛速度分别是什么？

$$\frac{1}{h^2}[(1+h)-e^h]$$

用上题中所给的形式表示极限．

10. 说明下列论断不正确．

a. 当 $x\to 0$ 时，$e^x-1=\mathcal{O}(x^2)$

b. 当 $x\to 0$ 时，$x^{-2}=\mathcal{O}(\cot x)$

c. 当 $x\to 0$ 时，$\cot x=o(x^{-1})$

11. 设$[a_n]\to 0,\lambda>1$．证明：当 $n\to\infty$ 时，$\sum_{k=0}^{n}a_k\lambda^k=o(\lambda^n)$．

12. 解释为什么最小上界公理不适用于空集.
13. 方程

$$(x^3-1)y+e^x y^2+\cos x-1=0$$

隐式地定义了两个函数，求这两个显函数.

25

14. 解微分方程时，常常得到隐式解. 证明方程

$$2x^3y^2+x^2y+e^x=c$$

定义了微分方程

$$\frac{dy}{dx}=-(6x^2y^2+2xy+e^x)/(4x^3y+x^2)$$

的一个解.
15. 天文学中的开普勒方程是 $x-y+\epsilon\sin y=0$，这里 ϵ 是范围 $0\leqslant\epsilon\leqslant1$ 中的一个参数. 证明对于每个实数 x，存在一个实数 y 使得上述等式成立. 证明若 $0\leqslant\epsilon<1$，则 dy/dx 处处连续. 提示：记 $x=y-\epsilon\sin y$ 并且分别考虑 $y-\epsilon\sin y$ 在 $y\to-\infty$ 和 $y\to+\infty$ 时的性态. 第二个问题可用隐函数定理.
16. 求点 x 使得在该点处方程

$$y-\ln(x+y)=0$$

隐式地定义了一个 x 的函数 y. 计算 dy/dx.
17. 举例说明为什么最小上界公理不适用于有理数集.
18. 最小上界公理是否适用于整数集?
19. 求下列式子的值.

   a. $\sup\limits_{x\in\mathbb{R}} \arctan x$

   b. $\sup\limits_{x\geqslant0} e^{-x}$

   c. $\inf\limits_{x\in\mathbb{R}} e^{-x}$

   d. $\sup\limits_{x\in\mathbb{R}} (x^2+1)^{-1}$
20. 用积分中值定理证明：对某个 $y\in(0,\pi/2)$，使得

$$\int_0^{\pi/2} e^x\cos x\,dx=e^y$$

21. 举例说明为什么 u 的连续性不能从定理 1 的条件中去掉.
22. 证明：若 $0<\theta<1$，则 $(1+a\theta^n)/(1+a\theta^{n-1})$ 线性收敛于 1.
23. 下列两式是否等价?

   a. 当 $|x|\to\infty$ 时，对某个 $\epsilon>0$，$|f(x)|=\mathcal{O}(|x|^{-n-\epsilon})$

   b. $|f(x)|=o(|x|^{-n})$
24. 证明$\mathbb{R}$中集合 S 的所有上界组成的集合或者是$\mathbb{R}$，或者是空集合，或者是形如$[a,\infty)$的区间.
25. 用归纳法证明霍纳算法是正确的.
26. 当计算序列 $x_n=(1+1/n)^n$ 时，发现它似乎是单调递增的. 证明它的确如此. 提示：首先，若 $\ln f(x)$ 递增，则 $f(x)$ 也递增. 其次，若 $f'(x)>0$，则 f 递增. 最后，定义 $\ln x=\int_1^x t^{-1}dt$.
27. (续)证明上题中序列的元素恒小于 3.
28. 证明：$x_n=x+o(1)$ 当且仅当 $\lim\limits_{x\to\infty}x_n=x$.

26

29. 证明课本中(1)～(5)式每个表达式的正确性.
30. 对于固定的 n，证明当 $x\to0$ 时，有

$$\sum_{k=0}^{n} x^k = 1/(1-x) + o(x^n)$$

31. 求 β 的最佳整数值使得对固定的 n，当 $x \to 0$ 时，有

$$\frac{1}{1-x} = 1 + x + x^2 + \cdots + x^n + \mathcal{O}(x^\beta) \quad (0 < x < 1)$$

对 $o(x^\beta)$，重复求一次．在这种情况下是否存在一个最佳整数值？为什么？

32. 证明：若 $x_n = \mathcal{O}(\alpha_n)$，则 $cx_n = \mathcal{O}(\alpha_n)$.
33. 证明：若 $x_n = \mathcal{O}(\alpha_n)$，则 $x_n/\ln n = o(\alpha_n)$.
34. 求 k 的最佳整数值使得当 $x \to 0$ 时，$\cos x - 1 + x^2/2 = \mathcal{O}(x^k)$.
35. 证明：若 $x_n = o(\alpha_n)$，则 $x_n = \mathcal{O}(\alpha_n)$．说明反之不成立.
36. 证明：若 $x_n = \mathcal{O}(\alpha_n)$，$y_n = \mathcal{O}(\alpha_n)$，则 $x_n + y_n = \mathcal{O}(\alpha_n)$.
37. 证明：若 $x_n = o(\alpha_n)$，$y_n = o(\alpha_n)$，则 $x_n + y_n = o(\alpha_n)$.
38. 证明：对于任意 $r > 0$，当 $x \to \infty$ 时，$x^r = \mathcal{O}(e^x)$.
39. 证明：对于任意 $r > 0$，当 $x \to \infty$ 时，$\ln x = \mathcal{O}(x^r)$.
40. 证明：若 $\alpha_n \to 0$，$x_n = \mathcal{O}(\alpha_n)$，并且 $y_n = \mathcal{O}(\alpha_n)$，则 $x_n y_n = o(\alpha_n)$.
41. 证明：若 $x_n = \mathcal{O}(\alpha_n)$，则 $\alpha_n^{-1} = \mathcal{O}(x_n^{-1})$．并证明对于 o 关系，同样成立.

计算机习题 1.2

1. 考虑递归关系

$$\begin{cases} x_0 = 1 \quad x_1 = c \\ x_{n+1} = x_n + x_{n-1} \quad (n \geqslant 1) \end{cases}$$

a. 证明：当 $c = (1+\sqrt{5})/2$ 时，

$$x_n = \left(\frac{1+\sqrt{5}}{2}\right)^n$$

给出一个闭型公式.

b. 类似证明：当 $c = (1-\sqrt{5})/2$ 时，

$$x_n = \left(\frac{1-\sqrt{5}}{2}\right)^n$$

给出一个闭型公式.

c. 若 $c = 1$，则

$$x_n = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^{n+1}$$

对于范围 $1 \leqslant n \leqslant 30$ 中的所有 n，用递归关系以及每种情况的公式同时计算 x_n．并解释计算结果．这个递归式定义了著名的**斐波那契序列**.

2. 用符号操作程序，求函数 $(\tan x)^2$ 在 0 处的泰勒级数，其中级数达到并且包含 x^{10} 项．用 $\mathcal{O}$ 记号表示略去的项.

[27] 3. 建立(1)～(5)式.

1.3 差分方程

数值计算的算法常常被设计成产生数的序列，因而讨论一些线性序列空间的理论是有益的．在第 8 章中将需要这个理论来分析线性多步法．（在这里提出它是因为理解它需要少量的

数学背景，从而适合独立成节给予介绍.）

1.3.1 基本概念

在下面讨论中，V 表示所有诸如

$$x=[x_1,x_2,x_3,\cdots]$$
$$y=[y_1,y_2,y_3,\cdots]$$

那样的无穷复数序列组成的集合. 形式上，序列是一个定义在正整数N={1, 2, 3, …}上的复值函数. 我们用 x_n 代替函数 x 在自变量 n 处的值 $x(n)$ 仅仅是为了方便.

在集合 V 中，我们定义两个运算：

$$x+y=[x_1+y_1,x_2+y_2,x_3+y_3,\cdots]$$
$$\lambda x=[\lambda x_1,\lambda x_2,\lambda x_3,\cdots]$$

表示这两个运算更紧凑些的方法是：

$$(x+y)_n=x_n+y_n$$
$$(\lambda x)_n=\lambda x_n$$

在 V 中，存在一个 0 元：$0=[0,0,0,\cdots]$. 采用这些定义后，V 就成为一个向量空间. 逐条验证向量空间是很麻烦的且不具有启发性. 向量空间 V 是无限维的. 确实，下面一组向量是线性无关的：

$$v^{(1)}=[1,0,0,0,\cdots]$$
$$v^{(2)}=[0,1,0,0,\cdots]$$
$$v^{(3)}=[0,0,1,0,\cdots]$$
$$v^{(4)}=[0,0,0,1,\cdots]$$
$$\vdots$$

我们将关注线性算子 $L: V\to V$. 在这类算子中最重要的一个是**移位算子**或**位移算子**，记作 E，它是由等式

$$Ex=[x_2,x_3,x_4,\cdots],\text{其中 } x=[x_1,x_2,x_3,\cdots]$$

28

定义的. 因此

$$(Ex)_n=x_{n+1}$$

显然 E 能以连续乘积的形式多次应用，例如，

$$(EEx)_n=x_{n+2}$$

或

$$(E^k x)_n=x_{n+k}$$

下面，我们把注意力放在能用 E 的幂的线性组合来表示的线性算子上. 这种算子称为(具有常系数和有限秩的)**线性差分算子**. 它的一般形式是

$$L=\sum_{i=0}^{m}c_iE^i \tag{1}$$

当然，E^0 定义为恒等算子，

$$(E^0x)_n=(Ix)_n=x_n$$

从(1)式，可看到它的线性差分算子构成一个从 V 到 V 的所有线性算子所组成集合的一个线性子空间. E 的所有幂构成这个子空间的一个基.

注意到(1)式中的 L 是 E 的一个多项式；换句话说，它是 E 的幂的线性组合. 这样，我们可以记

$$L = p(E)$$

其中 p 是一个多项式，称为 L 的**特征多项式**，并且定义为

$$p(\lambda) = \sum_{i=0}^{m} c_i \lambda^i$$

这里研究的是确定形如 $Lx=0$ 的方程的所有解，其中 L 是(1)型的算子. 从 L 的线性性，立刻得到集合 $\{x: Lx=0\}$ 是 V 的线性子空间；称之为 L 的**零空间**. 如果可求出 L 零空间的一个基，则方程 $Lx=0$ 可解.

下面我们考虑 L 的一个实例，譬如取 $c_0=2$，$c_1=-3$，$c_2=1$，其他 $c_i=0$. 这样得到的方程，称之为**线性差分方程**. 它能写成下面三种形式：

$$\begin{gathered} (E^2 - 3E^1 + 2E^0)x = 0 \\ x_{n+2} - 3x_{n+1} + 2x_n = 0 \quad (n \geqslant 1) \\ p(E)x = 0 \quad p(\lambda) = \lambda^2 - 3\lambda + 2 \end{gathered} \tag{2}$$

29

非常容易产生(2)解的序列. 实际上，可任意选择 x_1 和 x_2，再用(2)来确定 x_3，x_4，…. 例如，我们可用这种方法得到许多解

$$[1,0,-2,-6,-14,-30,\cdots]$$
$$[1,1,1,1,\cdots]$$
$$[2,4,8,16,\cdots]$$

第一个解要比后两个解难以理解，因为直观上看不出其通项是什么样的. 而后两个解显然具有形式 $x_n=\lambda^n$，$\lambda=1$ 或 2. 这样自然就要问是否存在任何其他这类解. 把 $x_n=\lambda^n$ 代入(2)得

$$\lambda^{n+2} - 3\lambda^{n+1} + 2\lambda^n = 0$$
$$\lambda^n(\lambda^2 - 3\lambda + 2) = 0$$
$$\lambda^n(\lambda - 1)(\lambda - 2) = 0$$

这个简单的分析表明具有所寻形式的其他解只有一个，即[0, 0, 0, …]，我们称这个解为**平凡解**. 现在发现由 $u_n=1$ 定义的解 u 和由 $v_n=2^n$ 定义的解 v 形成了(2)的解空间的一个基. 为证明这个结论，设 x 是(2)的**任意解**，寻找常数 α 和 β 以使得 $x=\alpha u+\beta v$. 这个等式意味着对所有的 n，有 $x_n=\alpha u_n+\beta v_n$. 特别，对 $n=1$，2，我们有

$$\begin{cases} x_1 = \alpha + 2\beta \\ x_2 = \alpha + 4\beta \end{cases} \tag{3}$$

因为矩阵

$$\begin{bmatrix} 1 & 2 \\ 1 & 4 \end{bmatrix}$$

的行列式不等于 0(这是一个在 6.1 节中定义的**范德蒙德矩阵**的行列式)，所以方程(3)唯一确定了 α 和 β. 现在用归纳法来证明对所有的 n，$x_n=\alpha u_n+\beta v_n$. 确实，如果指标小于 n 时等式成

立，那么对于 n 这个等式也应成立，因为

$$\begin{aligned} x_n &= 3x_{n-1} - 2x_{n-2} \\ &= 3(\alpha u_{n-1} + \beta v_{n-1}) - 2(\alpha u_{n-2} + \beta v_{n-2}) \\ &= \alpha(3u_{n-1} - 2u_{n-2}) + \beta(3v_{n-1} - 2v_{n-2}) \\ &= \alpha u_n + \beta v_n \end{aligned}$$

这例子说明了特征多项式单根[○]的情况.

30

1.3.2 单根

定理 1（零空间定理） 若 p 是一个多项式，λ 是 p 的一个根，则 $[\lambda, \lambda^2, \lambda^3 \cdots]$ 是差分方程 $p(E)x=0$ 的一个解. 若 p 的所有根是非零单根，则差分方程的每个解是这些特解的一个线性组合.

证明 首先，若 λ 是任意的复数，$u=[\lambda, \lambda^2, \lambda^3, \cdots]$，则因为

$$(Eu)_n = u_{n+1} = \lambda^{n+1} = \lambda(\lambda^n) = \lambda u_n$$

所以 $Eu = \lambda u$. 再运用算子 E，一般可以得到 $E^i u = \lambda^i u$. 因为 E^0 被定义为恒等映射，所以 $E^0 u = \lambda^0 u$. 因而若 p 是由 $p(\lambda) = \sum_{i=0}^{m} c_i \lambda^i$ 定义的多项式，则

$$p(E)u = \left(\sum_{i=0}^{m} c_i E^i\right)u = \sum_{i=0}^{m} c_i (E^i u) = \sum_{i=0}^{m} c_i \lambda^i u = p(\lambda)u$$

若 $p(\lambda)=0$，则如所断言的，$p(E)u=0$.

设 p 是一个多项式，其所有的根：$\lambda_1, \lambda_2, \lambda_3, \cdots, \lambda_m$ 是非零单根. 则对应于每个 λ_k，差分方程

$$p(E)x = 0$$

存在一个解. 即，有解 $u^{(k)} = [\lambda_k, \lambda_k^2, \lambda_k^3, \cdots]$. 设 x 是差分方程的任意一个解. 把它表示成 $x = \sum_{k=1}^{m} a_k u^{(k)}$. 取这级数的前 m 项，得到

$$x_i = \sum_{k=1}^{m} a_k \lambda_k^i \quad (1 \leqslant i \leqslant m) \tag{4}$$

具有元素 λ_k^i 的 $m \times m$ 矩阵是非奇异的，由于它的奇异性将推出非平凡的等式

$$\sum_{i=1}^{m} b_i \lambda_k^i = 0 \quad 或 \quad \sum_{i=1}^{m} b_i \lambda_k^{i-1} = 0$$

（最后的等式显示了一个有 m 个根的 $m-1$ 次多项式.）因而(4)式唯一决定了 $a_1, a_2, \cdots, a_m$. 接下来证明(4)式对所有的 i 值均成立. 令 $z = x - \sum_{k=1}^{m} a_k u^{(k)}$. 那么 $p(E)z = 0$ 或等价地对所有 n，$\sum_{i=0}^{m} c_i z_{n+i} = 0$. 换句话说，有

○ 当考虑多项式时，我们用术语根，对于其他(更一般的)函数，我们用术语零点. 在高等数学中，人们总是讲函数或多项式的零点，因为根是依照了较旧的用法.

$$z_{n+m} = -c_m^{-1}(c_0 z_n + c_1 z_{n+1} + \cdots + c_{m-1} z_{n+m-1}) \quad (n \geqslant 1) \tag{5}$$

由于多项式 p 有 m 个不同的根，因此其次数为 m，所以 $c_m \neq 0$. 因为 $z_i=0$，$i=1, 2, \cdots, m$，反复使用(5)式就得出

$$z_{m+1} = z_{m+2} = \cdots = 0 \tag{6}$$

31 ■

1.3.3 重根

此外还存在当 p 有重根时，求解差分方程 $p(E)x=0$ 的问题. 定义 $x(\lambda)=[\lambda, \lambda^2, \lambda^3, \cdots]$. 若 p 是任意多项式，我们已经看到

$$p(E)x(\lambda) = p(\lambda)x(\lambda) \tag{7}$$

对 λ 求导得

$$p(E)x'(\lambda) = p'(\lambda)x(\lambda) + p(\lambda)x'(\lambda)$$

若 λ 是 p 的重根，则 $p(\lambda)=p'(\lambda)=0$，(6)式和(7)式说明 $x(\lambda)$ 和 $x'(\lambda)$ 是此差分方程的解. 从而序列 $x'(\lambda)=[1, 2\lambda, 3\lambda^2, \cdots]$ 也是一个解. 若 $\lambda \neq 0$，则由于

$$\det\begin{bmatrix} \lambda & \lambda^2 \\ 1 & 2\lambda \end{bmatrix} \neq 0$$

因此它与解 $x(\lambda)$ 无关. 那么，当这个序列在第 2 项被截断时，得到 $\mathbb{R}^2$ 中的向量对是线性无关的.

通过扩展这种推理，可以证明，若 λ 是 p 的一个 k 重根，则下面的序列是差分方程 $p(E)x=0$ 的解：

$$\begin{aligned} x(\lambda) &= [\lambda, \lambda^2, \lambda^3, \cdots] \\ x'(\lambda) &= [1, 2\lambda, 3\lambda^2, \cdots] \\ x''(\lambda) &= [0, 2, 6\lambda, \cdots] \\ &\vdots \\ x^{(k-1)}(\lambda) &= \frac{\mathrm{d}^{(k-1)}}{\mathrm{d}\lambda^{k-1}}[\lambda, \lambda^2, \lambda^3, \cdots] \end{aligned}$$

定理 2（零空间的基定理） 设 p 是一个多项式，并且 $p(0)\neq 0$. 则可得到 $p(E)$ 的零空间的一个基如下：对于 p 的每个 k 重根 λ，有相应的 k 个基解 $x(\lambda)$，$x'(\lambda)$，$\cdots$，$x^{(k-1)}(\lambda)$，这里 $x(\lambda)=[\lambda, \lambda^2, \lambda^3, \cdots]$.

例 1 求差分方程

$$4x_n + 7x_{n-1} + 2x_{n-2} - x_{n-3} = 0$$

的通解.

解 已知方程具有 $p(E)x=0$ 形式，这里 $p(\lambda)=4\lambda^3+7\lambda^2+2\lambda-1$. p 的因式是 $(\lambda+1)^2$ 和 $(4\lambda-1)$. 因此，p 有 2 重根 -1 和单根 $1/4$. 所以，基解是

$$\begin{aligned} x(-1) &= [-1, 1, -1, 1, \cdots] \\ x'(-1) &= [1, -2, 3, -4, \cdots] \\ x\left(\frac{1}{4}\right) &= \left[\frac{1}{4}, \frac{1}{16}, \frac{1}{64}, \cdots\right] \end{aligned}$$

32

通解是

$$x = \alpha x(-1) + \beta x'(-1) + \gamma x\left(\frac{1}{4}\right)$$

或

$$x_n = \alpha(-1)^n + \beta n(-1)^{n-1} + \gamma\left(\frac{1}{4}\right)^n n$$

■

1.3.4 稳定的差分方程

如果对于 V 的元素 $x=[x_1, x_2, \cdots]$，存在一个常数 c 使得对所有的 n，有 $|x_n| \leqslant c$，换句话说，$\sup_n |x_n| < \infty$，则称 x **有界**. 若形如 $p(E)x=0$ 的差分方程的解有界，则称此差分方程是**稳定的**. 因为差分方程(2)的解之一为 $x_n=2^n$，所以它是不稳定的.（别处所讨论的条件和稳定性与此处稳定的差分方程这个概念无关.）

那么是否存在一个简单的方法可用来识别稳定的差分方程呢?

定理 3(稳定的差分方程定理)　*对于一个满足 $p(0)\neq 0$ 的多项式 p，下面这些条件是等价的：*

1. 差分方程 $p(E)x=0$ 是稳定的.

2. p 的所有根满足 $|z| \leqslant 1$，并且所有重根满足 $|z| < 1$.

证明　假定性质 2 对于 p 成立. 设 λ 是 p 的一个根，则 $x(\lambda)=[\lambda, \lambda^2, \lambda^3, \cdots]$是 p 相应的差分方程的一个解. 因为 $|\lambda| \leqslant 1$，所以这个序列有界. 若 λ 是重根，则 $x'(\lambda)$，$x''(\lambda)$，…中至少有一个也是差分方程的解. 此时，由性质 2，得 $|\lambda| < 1$. 根据初等微积分(洛必达法则)，得

$$\lim_{n\to\infty} n^k \lambda^n = 0 \quad (k \geqslant 0)$$

所以每个序列 $x'(\lambda)$，$x''(\lambda)$，…有界(见习题 1.3.22～1.3.23).

反之，假设性质 2 对于 p 不成立. 若 p 有一个根 λ 满足 $|\lambda| > 1$，则序列 $x(\lambda)$无界. 若 p 有一个重根 λ 满足 $|\lambda| \geqslant 1$，则序列 $x'(\lambda)$无界，因为它的通项满足下面不等式

$$|x_n| = |n\lambda^{n-1}| = n|\lambda|^{n-1} \geqslant n$$

■

例 2　判定下列差分方程是否稳定.

$$4x_n + 7x_{n-1} + 2x_{n-2} - x_{n-3} = 0$$

解　已知方程具有 $p(E)x=0$ 形式，$p(\lambda)=4\lambda^3+7\lambda^2+2\lambda-1$. 根据前例，$p$ 有 2 重根 -1 和单根 1/4. 因此，该方程**不稳定**. ■

33

在贝塞尔函数理论中有一个非常数系数的差分方程的例子. 公式

$$J_n(x) = \frac{1}{\pi}\int_0^\pi \cos(x\sin\theta - n\theta)\mathrm{d}\theta$$

定义了**贝塞尔函数** J_n. 因此显然有 $|J_n(x)| \leqslant 1$. 递归公式

$$J_n(x) = 2(n-1)x^{-1}J_{n-1}(x) - J_{n-2}(x)$$

虽然也同样成立，但是不直观. 若(对于某个 x)我们知道 $J_0(x)$和 $J_1(x)$，则能用递归关系来计算 $J_2(x)$，$J_3(x)$，…，$J_n(x)$. 由于必然出现的舍入误差将被乘上因式 $2nx^{-1}$，因此当 $2n > |x|$ 时，这个过程会变得不稳定和无用处. 而且该因式最终要变得非常大(见计算机习题

1.3.2).

有关利用递归关系计算函数的更多资料见 Abramowitz and Stegun[1964，第13页]、Cash[1979]、Gautschi[1961，1967，1975]以及 Wimp[1984].

习题 1.3

1. 对于方程(2)产生的序列，把其第1项表示成第2项和第3项的线性组合.
2. 设 p 是 m 次多项式. 试问 $p(E)x=0$ 的解空间是否一定为 m 维?
3. 设 p 是 m 次多项式，并且 $p(0)\neq 0$. 证明：若序列 x 包含连续 m 个 0 并且 $p(E)x=0$，则 $x=0$.
4. 算子 E 是否单射(一一对应)? 它是否有左逆或右逆? 它是否满射(映上)? 由 $(Fx)_n=x_{n-1}$，$(Fx)_1=0$ 定义算子 F，请对 F 回答同样的问题. 研究 E 和 F 之间的关系. 假设 V 重新定义为所有在集合{…，−3，−2，−1，0，1，2，3，…}上定义的函数所组成的空间，并且只由 $(Fx)_n=x_{n-1}$ 来定义算子 F. 回答前面的问题将会受到怎样的影响?
5. 算子 E 的特征值和特征向量是什么?
6. 考虑无穷级数 $\sum_{n=1}^{\infty} x_n v^{(n)}$. 关于其收敛性你能说明什么? 在逐点意义下，证明 $x=\sum_{n=1}^{\infty} x_n v^{(n)}$.
7. 当 $\{v^{(1)},v^{(2)},\cdots\}$ 是 V 的一个基时，证明 $\sum_{i=1}^{m} c_i E^i$ 可用一个无穷矩阵表示.
8. (续)证明任何具有上题所述形式的两个算子可以互相交换.
9. 证明：若 L_1 和 L_2 是 E 的幂的线性组合，并且 $L_1x=0$，则 $L_1L_2x=0$.
10. 对差分方程 $E^r x=0$，建立一个完整的理论.
11. 给出下列每个解空间中由实序列组成的基
    a. $(4E^0-3E^2+E^3)x=0$
    b. $(3E^0-2E+E^2)x=0$ [34]
    c. $(2E^6-9E^5+12E^4-4E^3)x=0$
    d. $(\pi E^2-\sqrt{2}E+\log 2\cdot E^0)x=0$
12. 证明：若 p 是一个实系数多项式，并且 $z\equiv[z_1,\ z_2,\ \cdots]$ 是 $p(E)z=0$ 的一个(复数)解，则 z 的共轭，z 的实部，z 的虚部也都是其解.
13. 解
    a. $x_{n+1}-nx_n=0$
    b. $x_{n+1}-x_n=n$
    c. $x_{n+1}-x_n=2$
14. 由
$$\Delta x=[x_2-x_1,x_3-x_2,x_4-x_3,\cdots]$$
来定义算子 Δ. 证明 $E=I+\Delta$. 并且证明：若 p 是一个多项式，则
$$p(E)=p(I)+p'(I)\Delta+\frac{1}{2}p''(I)\Delta^2+\frac{1}{3!}p'''(I)\Delta^3+\cdots+\frac{1}{m!}p^{(m)}(I)\Delta^m$$
15. (续)证明：若 $x=[\lambda,\ \lambda^2,\ \lambda^3,\ \cdots]$ 且 p 是一个多项式，则 $p(\Delta)x=p(\lambda-1)x$. 叙述如何求解写成 $p(\Delta)x=0$ 形式的差分方程.
16. (续)证明
$$\Delta^n=(-1)^n\left[E^0-nE+\frac{1}{2}n(n-1)E^2-\frac{1}{3!}n(n-1)(n-2)E^3+\cdots+(-1)^nE^n\right]$$

17. 给出定理 2 的完整证明.
18. 设 p 是一个多项式，并且满足 $p(0)=0$. 描述 $p(E)$的零空间.
19. 对 $\lambda\in\mathbb{C}$，定义 $x(\lambda)=[\lambda, \lambda^2, \lambda^3, \cdots]$. 证明：若 $\lambda_1, \lambda_2, \cdots, \lambda_m$ 是不同的非零复数，则$\{x(\lambda_1), x(\lambda_2), \cdots, x(\lambda_m)\}$是 V 中的线性无关集.
20. 证明：若 λ 是多项式 p 的非零 k 重根，则方程 $p(E)x=0$ 有解 $u^{(1)}, u^{(2)}, \cdots, u^{(k)}$，其中 $u_n^{(j)}=n^{j-1}\lambda^n$.
21. 证明：若 $\mu\in(0, \infty)$且 $|\lambda|<1$，则$\lim\limits_{n\to\infty}n^\mu\lambda^n=0$.
22. 详细证明收敛序列是有界的.
23. 在不假设 $p(0)\neq 0$ 的情况下证明定理 3.
24. 由等式 $x_{n+1}=x_n+x_n^{-1}$，其中 $x_0>0$，归纳定义一个序列. 确定当 $n\to\infty$时，x_n 的性态.
25. 确定差分方程 $x_n=x_{n-1}+x_{n-2}$是否稳定.
26. 证明：若 x 是差分方程 $p(E)x=0$ 的解，则 Ex 也是其解.
27. 考虑递归关系 $x_n=2(x_{n-1}+x_{n-2})$. 证明其通解是 $z_n=\alpha(1+\sqrt{3})^n+\beta(1-\sqrt{3})^n$. 并证明初始值为 $x_1=1$, $x_2=1-\sqrt{3}$的解对应于 $\alpha=0$, $\beta=(1-\sqrt{3})^{-1}$.

计算机习题 1.3

1. 考虑差分方程 $x_{n+2}-2x_{n+1}-2x_n=0$. $x_n=(1-\sqrt{3})^{n-1}$是它的一个解. 该解是符号振荡的序列并且收敛于 0. 用等式 $x_{n+2}=2(x_{n+1}+x_n)$，初始值为 $x_1=1$, $x_2=1-\sqrt{3}$，计算且打印出这个序列的前 100 项. 解释所出现的不寻常现象.

35

2. 以 $J_0(1)=0.765\ 197\ 686\ 6$, $J_1(1)=0.440\ 050\ 585\ 7$ 作为初始值，用课本中的递归公式计算 $J_2(1)$, $J_3(1), \cdots, J_{20}(1)$. 那么在总的误差值中存在什么现象?
3. 考虑差分方程 $4x_{n+2}-8x_{n+1}+3x_n=0$. 判断它是否稳定，并且求它的通解. 假定 $x_0=0$, $x_1=-2$，用最有效的方法计算 x_{100}.
4. 用下面 3 种方法计算习题 1.3.27 的特解. 对 $1\leqslant n\leqslant 100$，计算且比较这些解.
   a. x_n 直接由递归关系计算
   b. $y_n=\beta(1-\sqrt{3})^n$
   c. $z_n=\alpha(1+\sqrt{3})^n+\beta(1-\sqrt{3})^n$，这里选择 α 为计算机的单位舍入误差
5. 数值求解差分方程 $x_{n+2}-(\pi+\pi^{-1})x_{n+1}+x_n=0$, $x_0=1$, $x_1=\pi$. 计算 49 项并且求出在 x_{50} 中的相对误差. 再把 x_1 变成 π^{-1}，做同样的工作且解释在这两种情况中相对误差的区别.

36

第2章　计算机算术运算

2.0　概述

在本章中，我们叙述浮点数系并且阐明可能损害计算机计算的舍入误差的一些基本事实．我们还讨论其他类型的误差和有效位丢失．当两个几乎相等的数相减时，就会出现有效位丢失．最后，我们综述若干稳定/不稳定的算法和病态问题．

2.1　浮点数和舍入误差

与人类更喜欢使用十进制数系不同，许多计算机用二进制数系处理实数．与十进制以10为基数的方法一样，二进制以2为基数．为了作比较，首先回顾一下如何处理我们熟悉的数的表示．在**十进制**中把一个像427.325这样的实数更详细地写出时，我们有

$$427.325 = 4 \times 10^2 + 2 \times 10^1 + 7 \times 10^0 + 3 \times 10^{-1} + 2 \times 10^{-2} + 5 \times 10^{-3}$$

表达式的右边包含10的幂和数字0，1，2，3，4，5，6，7，8，9．若我们准许在小数点右边可以出现无穷多个数字，则任何实数都能以刚才说明的方式并在其前面加上符号（＋或－）来表示．因此，例如，$-\pi$是

$$-\pi = -3.141\ 592\ 653\ 589\ 793\ 238\ 462\ 643\ 38\cdots$$

写在最后面的8表示8×10^{-26}．

在二进制中，仅使用两个数字0和1．用**二进制**我们也能详细写出一个典型的数．例如，

$$\begin{aligned}(1001.11101)_2 &= 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 \\ &\quad + 1 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-3} + 0 \times 2^{-4} + 1 \times 2^{-5}\end{aligned}$$

这个数与用十进制记数法表示的实数9.906 25相同．（请验证．）

37

一般地，任一整数$\beta > 1$都可用作一个数系的基数．这些以β为基数来表示的数包含数字0，1，2，3，4，…，$(\beta - 1)$．若从上下文不能区别数N用怎样的基数，则可使用符号$(N)_\beta$．因此从上面的讨论，我们有

$$(1001.11101)_2 = (9.906\ 25)_{10}$$

因为计算机内部以**二进制**方式工作而以**十进制**方式与其用户进行交流，所以转换过程必须由计算机来执行．这些过程是在输入和输出时执行的．通常，用户不必关心这些转换，但是它们牵涉了一些小的舍入误差．

计算机不能对超过固定位数的实数进行运算．计算机字长对可表示实数的精度给予限制．因此，甚至像1/10这样简单的数也不能在任何二进制计算机中精确地存储，因为它需要一个无穷的二进制表达式：

$$\frac{1}{10} = (0.00011001100110011\cdots)_2 \tag{1}$$

例如，若我们把0.1读入一个32位的计算机工作站，然后把它的40位小数打印出来，则得到下面的结果：

0.100 000 001 490 116 119 384 765 625 000 000 000 000 0

通常，我们不会注意到这种转换误差，因为用默认格式打印出的是0.1.

此外，在使用计算机时，我们应该意识到这里涉及了两种转换技术——从十进制转换出和转换回十进制——因为我们更愿意用十进制进行计算，而计算机却愿意用二进制. 误差可能在每次转换中产生.

2.1.1 舍入

为什么现在要讨论舍入？在本章后面，我们将对舍入进行详细的讨论并且把它与计算机计算联系起来. 这里所提及的舍入仅仅与用手工计算或用袖珍计算器计算有关. 其原因是在科学计算中，中间结果中数字的个数可能变得越来越大，而有效位数保持不变或减少. 例如，两个小数点右边有8位数字的数之积是一个小数点右边有16位数字的数.

在科学计算中舍入是一个重要的概念. 考虑一个具有形式0.□□□…□□□的正的十进制小数x，其小数点右边有m位. 根据第$(n+1)$位的值，把x**舍入**到n位小数$(n<m)$. 若第$(n+1)$位的数字是0，1，2，3或4，则第n位的数字不变，并丢弃后面的所有数字. 若它是5，6，7，8或9，则第n位的数字增加一个单位并且丢弃后面剩余的数字.（第$(n+1)$位的数字是5的情况有多种处理方法. 例如，只有当前面数字是偶数时，选择上舍入，这差不多有一半机会出现. 为了方便，在第$(n+1)$位的数字是5的情况下，我们总是选择上舍入.）

38

下面是一些7位数被正确地舍入到4位数的例子：

$$0.173\,5 \leftarrow 0.173\,549\,9$$
$$1.000 \leftarrow 0.999\,950\,0$$
$$0.432\,2 \leftarrow 0.432\,160\,9$$

若x被舍入使得$\tilde{x}$是它的n位近似，则

$$|x-\tilde{x}| \leqslant \frac{1}{2}\times 10^{-n} \tag{2}$$

要知道这为什么是正确的，讨论如下：若x的第$(n+1)$位是0，1，2，3或4，则$x=\tilde{x}+\epsilon$，其中$\epsilon<\frac{1}{2}\times 10^{-n}$，因此不等式(2)成立. 若$x$的第$(n+1)$位是5，6，7，8或9，则$\tilde{x}=\hat{x}+10^{-n}$，这里$\hat{x}$的前$n$位数字与$x$相同，超过第$n$位的数字都是0. 现在$x=\hat{x}+\delta\times 10^{-n}$，其中$\delta\geqslant 1/2$并且$\tilde{x}-x=(1-\delta)\times 10^{-n}$. 因为$1-\delta\leqslant 1/2$，所以不等式(2)成立.

若x是十进制小数，则x的**切断或截断n**位近似就是通过直接丢弃所有超过第n位的数字而获得的数$\hat{x}$. 对这个数$\hat{x}$，我们有

$$|x-\hat{x}| < 10^{-n} \tag{3}$$

x和$\hat{x}$之间的关系使得$x-\hat{x}$的前n位数字为0并且$x=\hat{x}+\delta\times 10^{-n}$，其中$0\leqslant\delta<1$. 因此，$|x-\hat{x}|=|\delta|\times 10^{-n}<10^{-n}$，由此可得不等式(3).

2.1.2 规格化的科学记数法

在十进制中，任何实数都能用**规格化的科学记数法**表示. 这意味着移动小数点和补充10的相应幂次使所有数字都在小数点右边并且第1位数字不是0. 例如

$$732.505\ 1 = 0.732\ 505\ 1 \times 10^3$$
$$-0.005\ 612 = -0.561\ 2 \times 10^{-2}$$

一般而言，一个非零实数 x 可表示成

$$x = \pm r \times 10^n$$

形式，其中 $1/10 \leqslant r < 1$，n 为整数(正数、负数或零)．当然，若 $x=0$，则 $r=0$；在其他所有情况，我们可以调节 n 使 r 处于给定范围内．

39

用完全相同的方法，我们能在**二进制**中使用科学记数法．我们有

$$x = \pm q \times 2^m \tag{4}$$

其中 $\frac{1}{2} \leqslant q < 1$(如果 $x \neq 0$)，m 为整数．称数 q 为**尾数**，称整数 m 为**指数**．q 和 m 都是以 2 为基数的数．

当在计算机中存储二进制数时，最好对(4)式稍加修改．假设把二进制数首位数字 1 恰好移到二进制小数点的左边．在这种情况下，表达式是 $q=(1.f)_2$ 并且 $1 \leqslant q < 2$．现在仅把 $(.f)_2$ 以计算机字的形式存储，以便利用实际不存储二进制首位数字 1 而仅仅假定它的存在来节省一位空间．当然，在规格化形式中我们可以不存储首位的 1，这实际上是一回事．这在下一小节中会变得清楚．

2.1.3 假想计算机 Marc-32

在典型计算机内部，以刚才所描述的方法来表示数，但通过有效字长对 q 和 m 施加某些限制．为了说明这一点，我们考虑一台称为 Marc-32 的假想计算机．它的字长为 32 位(二进制位)，因此，它类似于许多个人电脑和工作站．

如图 2-1 所示，在假想的 32 位计算机 Marc-32 中，单精度实数的浮点表示被分成 3 个字段．

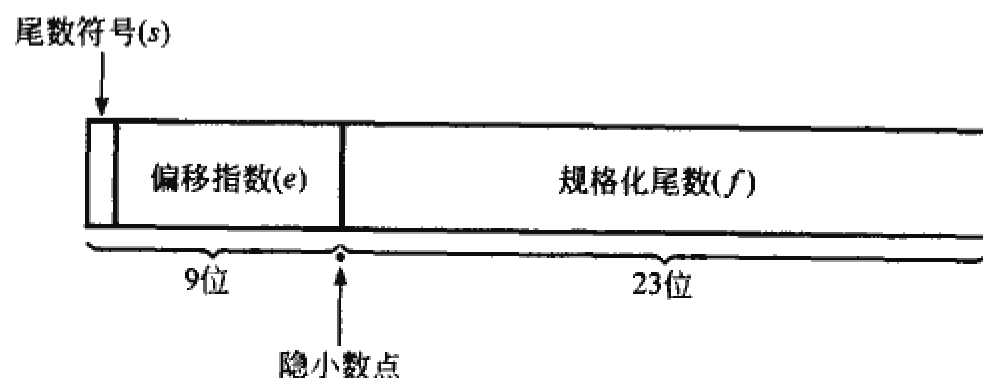


图 2-1 Marc-32 单精度字段

在 Marc-32 中，在表示一个非零实数 $x = \pm q \times 2^m$ 时用下面的方式分配组成一个字的二进制位：

实数 x 的符号	1 位
偏移指数(整数 e)	8 位
尾数部分(实数 f)	23 位

一个非零实数 $x=\pm q\times 2^m$ 能写成**左移规格化二进制数**使得尾数中第一个非零二进制数字恰好在二进制小数点前面，即 $q=(1.f)_2$. 这个非零二进制数字可假定是1且不需要存储. 尾数的范围是 $1\leqslant q<2$. 在这个字中保存尾数的23位用于存储来自 f 的23个二进制数字. 所以实际上，这台计算机的浮点数有24位尾数.

40

因此，非零规格化机器数是一个位串，其数值被解码如下：

$$x=(-1)^s q\times 2^m \tag{5}$$

其中

$$q=(1.f)_2 \text{并且} m=e-127$$

这里 $1\leqslant q<2$ 并且 q 中最高有效数字是1而且不显式存储. 此外，s 是表示 x 符号的位(正数用0表示，负数用1表示)，$m=e-127$ 是8位偏移指数，而 f 是实数 x 的23位小数部分，其与隐含的首位1一起得到有效数字字段$(1.\square\square\square\cdots\square\square\square)_2$.

一个用(5)式来表示的实数称为具有**规格化的浮点形式**. 因而，若它能用占8位的 $|m|$ 和占23位的 q 表示，则它是Marc-32中的一个**机器数**. 即它能在这个特殊的计算机中被精确表示. 大多数实数在Marc-32内不能被精确表示. 要是像这样的数作为输入数据或作为计算结果出现，那么在用**机器数**尽可能精确地表示它时就会产生不可避免的误差.

$|m|$ 表示不超过8位的限制意味着

$$0<e<(11111111)_2=2^8-1=255$$

并且值 $e=0$ 和 $e=255$ 为像 ±0，$\pm\infty$ 和NaN(非数字)这样的特殊情况所保留的. 因为 $m=e-127$，所以我们取 $-126\leqslant m\leqslant 127$ 并且Marc-32能处理小到与 $2^{-126}\approx1.2\times10^{-38}$ 一样小和大到与 $(2-2^{-23})2^{127}\approx3.4\times10^{38}$ 一样大的数. 对某些科学计算而言这并非是足够大的数量范围，而且因为这样或那样的一些原因，我们有时候必须用**双精度或扩充精度算术运算**编写程序. 双精度的浮点数用两个计算机字表示，并且尾数通常至少是原来位的两倍. 因此，双精度的精确小数位数大约是单精度的2倍. 用双精度计算比用单精度慢许多，常常是2倍或2倍以上. 这就是为什么通常双精度算术运算用软件完成而单精度算术运算用硬件完成的原因.

尾数部分要求不超过23位的限制意味着我们的机器数有大约6位十进制小数的有限精度，因为尾数的最低有效位表示 2^{-23} 个单位(或约 1.2×10^{-7}). 因此把超过6位十进制小数表示的数输入计算机，它们将是近似的. 此外，一些像1/100这样的简单的十进制数不是二进制计算机的机器数!

在二进制计算机中，浮点数分布相当不均匀，大部分聚集在0附近. (事实上，在实轴上存在许多间隙或孔. 例如，在0邻近的间隙常被称为**在零点的孔**.) 在计算机中仅仅存在有限个浮点数，而在2的毗连幂之间总是存在相同个数的机器数. 因为2的幂之间的间隙在0附近较小，而在远离0处较大，这产生了浮点数不均匀的分布，在原点附近密度较高.

41

除了必须用于存储符号的一个位外，**整数**能把所有的计算机字用于它的表示. 所以在Marc-32中，整数的范围是从 $-(2^{31}-1)$ 到 $2^{31}-1=2\ 147\ 483\ 647$. 在科学计算中，纯粹的整数计算并不常见.

2.1.4 零，无穷大，非数字

在IEEE标准算术运算中，单精度的0存在两种形式，-0和$+0$，计算机中分别用字$[00000000]_{16}$和$[80000000]_{16}$表示．大多数导致0值的算术运算给的是值$+0$．极其微小的并且对机器精度而言是0的负数给的是值-0．

类似地，单精度的无穷大存在两种形式，$-\infty$和$+\infty$，分别用计算机字$[7F800000]_{16}$和$[FF800000]_{16}$表示．通常，只要无穷大进行运算有意义，它就会被视作一个非常大的数．例如，假设x是浮点数并且$0<x<\infty$，于是计算$x+\infty$，$x*\infty$，∞/x中的每个结果都是值$+\infty$，而x/∞是$+0$．这里∞被理解为是$+\infty$．对$-\infty$有类似的结果．

NaN的意思是**非数字**并且产生于诸如$0/0$，$\infty-\infty$，$x+\text{NaN}$等不定型运算．用$e=255$和$f\neq0$的计算机字表示所有NaN．

2.1.5 机器舍入

除了对输入数据进行舍入外，大多数算术运算后也都需要进行舍入．算术运算结果驻留在长度为80位的计算机寄存器中，并且在放入到存储器之前必须将它舍入成单精度．在双精度运算中也出现类似的情形．

通常(默认)的舍入模式是**舍入到最接近数**：选择实数左右两边较近的那个机器数．在距离相同时，采用**舍入到偶数**：若实数正好在其左右两个机器数的中间，则选择偶机器数．用默认舍入模式(舍入到最接近数加上舍入到偶数)，最大误差是最低有效位上的半个单位．

舍入的其他模式包括**直接舍入**，例如**向0舍入**(也称**截断**)，**向$+\infty$舍入**，**向$-\infty$舍入**．

2.1.6 Fortran 90的内部过程

在Fortran 90中，大量的类属内部过程可以用来确定正在使用的计算机的数值环境．一般地，它们返回与自变量的**类型**(实型，整型，复型等)和**类别**(单精度、双精度等)相同的数．例如，一些与浮点数有关的内部过程：digits是有效(二进制)位数，epsilon是一个与单位相比几乎可以忽略的正数(使得$\epsilon+1\neq1$的最小浮点数ϵ)，huge是最大数，maxexponent是最大(二进制的)指数，minexponent是最小(最负二进制的)指数，precision是十进制精度，radix是计算机浮点数系的基数，range是十进制指数的范围，而tiny是最小正数．

42

表2-1是按在带IEEE标准算术运算的32位字的工作站上调用这些过程得到的结果编制的．对于整数x，digits(x)是31而huge(x)是$2\ 147\ 483\ 647\approx2^{31}-1$．

表2-1 内部过程的结果(单精度)

digits(x)	24	precision(x)	6
epsilon(x)	$1.192\ 092\ 9\text{E}-7\approx2^{-23}$	radix(x)	2
huge(x)	$3.402\ 823\ 4\text{E}+38\approx2^{128}$	range(x)	37
maxexponent(x)	128	tiny(x)	$1.175\ 494\ 4\text{E}-38\approx2^{-126}$
minexponent(x)	-125		

表 2-2 是根据在带 IEEE 标准算术运算的 32 位工作站(双精度)和 64 位字的 Cray T3E 超级计算机(单精度)上调用这些过程得到的结果编制的. 对于超级计算机上的整数 x, digits(x)是 63 而 huge(x)是 9 223 372 036 854 775 807≈$2^{63}-1$.

表 2-2 内部过程的结果(双精度)

digits(x)	53	precision(x)	15
epsilon(x)	2.220 446 049 250 313E-16 ≈ 2^{-52}	radix(x)	2
huge(x)	1.797 693 134 862 315 7E+308 ≈ $2^{1\,024}$	range(x)	307
maxexponent(x)	1 024	tiny(x)	2.225 073 858 507 201 3E-308 ≈ $2^{-1\,022}$
minexponent(x)	-1 021		

2.1.7 IEEE 标准浮点算术运算

Marc-32 的实数表示仿效 32 位计算机中通常的浮点表示, 它是 **IEEE 标准浮点表示**. 我们在这里仅给一个简短的描述. 例如, 根据目前正式的标准执行浮点算术运算的计算机对内部计算使用 80 位. 这里用到许多其他的概念——**保护位, 舍入位, 保留位, 不可规格化数, 非规格化数, 双重舍入**, 等等——这些概念涉及对这个主题的任何详细讨论. 限于篇幅, 在这里我们既不定义也不讨论它们, 但是我们介绍感兴趣的读者参考下面文献中的补充资料:《Standard for Binary Floating-Point Arithmetic》ANSI/IEEE[1985]和《A Radix-Independent Standard for Floating-Point Arithmetic》ANSI/IEEE[1987]. 关于这个主题有影响的论文有 Cody[1988]、Cody et al. [1984]、Coonen[1981]、Fosdick[1993]、Hough[1981]、Raimi[1969]和 Scott[1985].

43

2.1.8 接近的机器数

我们现在想要估计用 Marc-32 中一个接近的机器数作为一个给定正实数 x 的近似值所涉及的误差. 假定

$$x = q \times 2^m, \quad 1 \leqslant q < 2, \quad -126 \leqslant m \leqslant 127$$

我们要问, 什么机器数最接近于 x? 首先, 记

$$x = (1.a_1 a_2 \cdots a_{23} a_{24} a_{25} \cdots)_2 \times 2^m$$

其中每个 a_i 或者是 0 或者是 1. 通过简单地丢弃超出位 $a_{24}a_{25}\cdots$得到一个接近的机器数. 这个过程通常称为**截断**. 所得到的数是

$$x_- = (1.a_1 a_2 \cdots a_{23})_2 \times 2^m$$

注意到 x_- 在实轴上位于 x 的左边. 另一个接近的机器数位于 x 的右边. 它通过**上舍入**得到; 也就是说, 如前所述我们删去超出位但是在最后保留位 a_{23}上加一个单位. 这个数是

$$x_+ = ((1.a_1 a_2 \cdots a_{23})_2 + 2^{-23}) \times 2^m$$

如图 2-2 所示, 有两种情况. 在计算机中选择 x_- 和 x_+ 中较近的一个表示 x.

若 x 用 x_- 表示更好, 则

$$|x - x_-| \leqslant \frac{1}{2}|x_+ - x_-| = \frac{1}{2} \times 2^{m-23} = 2^{m-24}$$

图 2-2 x 的两种典型的位置

在这种情况中，**相对误差**有如下的界：

$$\left|\frac{x-x_-}{x}\right| \leqslant \frac{2^{m-24}}{q\times 2^m} = \frac{1}{q}\times 2^{-24} \leqslant 2^{-24}$$

44

在第二种情况中，x_+ 比 x_- 更接近 x，并且有

$$|x-x_+| \leqslant \frac{1}{2}|x_+-x_-| = 2^{m-24}$$

再作同样的分析，表明相对误差不大于 2^{-24}.（另一种界适用于**截断**过程，见习题 2.1.1.）

在计算过程中，有时产生形式为 $\pm q\times 2^m$ 的数，其中 m 超出计算机所允许的范围. 若 m 太大，则我们说发生**上溢**. 若 m 太小，就说发生**下溢**. 在 Marc-32 中，这意味着分别为 $m>127$ 和 $m<-126$. 在第一种情况(上溢)，存在致命的错误状态. 在许多计算机中，当含有 NaN 或机器无穷大的变量用于无意义的计算时，程序的执行自动停止. 在第二种情况(下溢)，许多计算机简单地把变量置 0 并且允许继续进行计算.（发布一个消息警告用户发生了下溢.）

我们总结一下刚才谈到的涉及 Marc-32 的内容：若 x 是在这台机器范围内的非零实数，则最接近于 x 的机器数 x^* 满足不等式

$$\left|\frac{x-x^*}{x}\right| \leqslant 2^{-24}$$

设 $\delta=(x^*-x)/x$，我们能把不等式写成下列形式：

$$\mathrm{fl}(x) = x(1+\delta) \qquad |\delta| \leqslant 2^{-24} \tag{6}$$

符号 $\mathrm{fl}(x)$用来表示最接近 x 的浮点机器数 x^*. 前面不等式中的数 2^{-24} 称为 Marc-32 的**单位舍入误差**. 分析表明分配给尾数的位数直接关系到计算机的单位舍入误差并且决定了计算机算术运算的精度.（对 Marc-32，计算机的最小浮点数 ϵ 是单位舍入误差的两倍.）

我们对 Marc-32 所描述的结论，经适当的修改，也适合其他计算机. 若一台计算机用 β 作基数并且其浮点数的尾数有 n 位，则

$$\mathrm{fl}(x) = x(1+\delta) \quad |\delta| \leqslant \epsilon$$

其中，在适当的舍入情况下 $\epsilon=\frac{1}{2}\beta^{1-n}$，而在截断情况下 $\epsilon=\beta^{1-n}$. 数 ϵ 是单位舍入误差而且是计算机及其操作系统和计算模式(不管是单精度还是多精度)的一个特征.

由于计算机的字长、用于算术运算的基数以及采用的舍入类型等不同，现代计算机的单位舍入误差(ϵ)的值大相径庭. 字长的变化从大型科学计算机的 64 和 60 位到中型机的 32 和 36 位，再到某些个人电脑的 16 位. 可编程的计算器可能有更小的精确度. 许多计算机用二进制的算术运算，但是也用十六进制和八进制. 采用的舍入类型有**完全舍入**、**伪舍入**和**截断**；有些编译器允许用户设置一个开关语句以便在计算中选择舍入类型.

45

假设 $x=q\times 2^m$ 是一个正的非零**机器数**. 通过改变 q 中末位数字，我们获得在右边的下一个(较大的)机器数

$$x_r=(q+2^{-23})\times 2^m$$

和在左边的前一个(较小的)机器数

$$x_\ell=(q-2^{-23})\times 2^m$$

显然，我们有

$$x_r-x=x-x_\ell=2^{m-23}$$

所以，

$$\frac{x_r-x}{x}=\frac{x-x_\ell}{x}=\frac{1}{q}\times 2^{-23}$$

因为 $1\leqslant q<2$，所以有

$$2^{-24}<\frac{x_r-x}{x}=\frac{x-x_\ell}{x}\leqslant 2^{-23}$$

因此，在任何机器数 x 与其两侧机器数 x_ℓ 和 x_r 中的任意一个之间相对间距近似地为一个不变值，即 2^{-23}. 这个值正好是计算机表示的精度.

例 1 数 $x=2/3$ 的二进制形式是什么? 在 Marc-32 中，两个接近的机器数 x_- 和 x_+ 是多少? 其中哪个作为 $\mathrm{fl}(x)$? 用 $\mathrm{fl}(x)$表示 x 时的绝对舍入误差和相对舍入误差各是多少?

解 为了确定二进制表示，我们记

$$\frac{2}{3}=(0.a_1a_2a_3\cdots)_2$$

用 2 乘得到

$$\frac{4}{3}=(a_1.a_2a_3\cdots)_2$$

所以，取两边的整数部分，得到 $a_1=1$. 两边减 1，我们有

$$\frac{1}{3}=(0.a_2a_3a_4\cdots)_2$$

重复前面步骤，最终获得

46

$$x=\frac{2}{3}=(0.101\ 0\cdots)_2=(1.010\ 101\cdots)_2\times 2^{-1}$$

两个接近的机器数是

$$x_-=(1.010\ 1\cdots 010)_2\times 2^{-1}$$
$$x_+=(1.010\ 1\cdots 011)_2\times 2^{-1}$$

这里 x_- 由截断得到，而 x_+ 由上舍入得到. 二进制小数点的右边有 23 位.

为了确定哪个更接近 x，我们计算 $x-x_-$ 和 x_+-x，从而决定从 x_- 和 x_+ 中应取哪个作为 $\mathrm{fl}(x)$:

$$x-x_-=(0.101\ 0\cdots)_2\times 2^{-24}=\frac{2}{3}\times 2^{-24}$$

$$x_+-x=(x_+-x_-)-(x-x_-)=2^{-24}-\frac{2}{3}\times 2^{-24}=\frac{1}{3}\times 2^{-24}$$

所以，取 $fl(x)=x_+$，而绝对舍入误差是

$$|fl(x)-x| = \frac{1}{3}\times 2^{-24}$$

于是，相对舍入误差是

$$\frac{|fl(x)-x|}{|x|} = \frac{\frac{1}{3}\times 2^{-24}}{\frac{2}{3}} = 2^{-25}$$

确定机器数 x_- 和 x_+ 的内部表示是有意义的．由于指数是－1，我们求出 $e=(126)_{10}=(176)_8=(001111110)_2$．因而，内部表示是

$$x_- = [00111111001010101010101010101010]_2 = [3F2AAAAA]_{16}$$

$$x_+ = [00111111001010101010101010101011]_2 = [3F2AAAAB]_{16}$$

当打印时，输出的十进制数是

$$x_- = 0.666\ 666\ 626\ 930\ 236\ 816\ 406\ 250\ 000\ 0$$

$$x_+ = 0.666\ 666\ 686\ 534\ 881\ 591\ 796\ 875\ 000\ 0$$

这里 $0.000\ 000\ 059\ 604\ 644\ 775\ 390\ 625\ 000\ 0=2^{-24}$ 是它们之间的绝对间距． ■

2.1.9 浮点误差分析

在继续研究因有限字长直接导致的计算机计算误差的过程中，我们将用 Marc-32 作为模型．假定这台机器的设计是这样的：每当两个机器数进行运算时，首先正确地形成结合，然后对结合进行规格化和舍入，最后以计算机字的形式把它存储在存储器中．为了更清楚，用符号 $\odot$ 表示四种基本算术运算＋，－，*，÷中的任意一种运算．设 x 和 y 是机器数，当 $x\odot y$ 被计算和存储时，我们得到的最接近于 $x\odot y$ 的数实际上是以一个舍入到 $fl(x\odot y)$ 机器字的形式拟合 $x\odot y$，然后存储那个数．

47

例 2 为了说明这个过程，我们用一台十进制计算机，其浮点数系中使用 5 位十进制小数，求两个机器数

$$x = 0.314\ 26\times 10^3 \qquad y = 0.925\ 77\times 10^5$$

的加、减、乘和除的相对误差．

解 对中间结果用双倍长的累加器，我们有

$$x+y = 0.928\ 912\ 600\ 0\times 10^5$$

$$x-y = -0.922\ 627\ 400\ 0\times 10^5$$

$$x*y = 0.290\ 932\ 480\ 2\times 10^8$$

$$x\div y \approx 0.339\ 457\ 964\ 7\times 10^{-2}$$

有 5 位小数的计算机以舍入形式

$$fl(x+y) = 0.928\ 91\times 10^5$$

$$fl(x-y) = -0.922\ 63\times 10^5$$

$$\mathrm{fl}(x*y)=0.290\,93\times10^{8}$$

$$\mathrm{fl}(x\div y)=0.339\,46\times10^{-2}$$

存储它们．这些结果的相对误差分别是2.8×10^{-6}，2.8×10^{-6}，8.5×10^{-6}，6.0×10^{-6}——所有的相对误差都小于10^{-5}．■

在任何计算机中，最吸引人的事情是知道四则算术运算满足像下面的式子：

$$\mathrm{fl}(x\odot y)=[x\odot y](1+\delta)\qquad|\delta|\leqslant\varepsilon$$

对于通用计算机，我们假定这个等式成立并且取ε为那台计算机的**单位舍入**．在任何设计完美的计算机中，这种假定的确成立或如此接近成立，以致误差在舍入分析中可被忽略．

先前对 Marc-32 建立的(6)式

$$\mathrm{fl}(x)=x(1+\delta)\qquad|\delta|\leqslant2^{-24}$$

其中x是 Marc-32 范围内的任意实数．因此，若x和y是机器数，则我们有

$$\mathrm{fl}(x\odot y)=(x\odot y)(1+\delta)\qquad|\delta|\leqslant2^{-24}$$

从而，单位舍入2^{-24}给出有关任何单个基本算术运算*相对误差*的一个界．前面数值例子的检验表明为什么在每次算术运算中必须假设舍入误差．若x和y不一定是机器数，则相应的结果是

[48]

$$\mathrm{fl}(\mathrm{fl}(x)\odot\mathrm{fl}(y))=(x(1+\delta_1)\odot y(1+\delta_2))(1+\delta_3)\qquad|\delta_i|\leqslant2^{-24}$$

用刚才得到的基本结果，我们能分析混合算术运算．为了说明，假设x，y和z是 Marc-32 中的机器数，并且希望计算$x(y+z)$．因为与2^{-23}相比，$\delta_2\delta_1$可以被忽略，所以有

$$\begin{aligned}\mathrm{fl}[x(y+z)]&=[x\mathrm{fl}(y+z)](1+\delta_1) &&|\delta_1|\leqslant2^{-24}\\&=[x(y+z)(1+\delta_2)](1+\delta_1) &&|\delta_2|\leqslant2^{-24}\\&=x(y+z)(1+\delta_2+\delta_1+\delta_2\delta_1)\\&\approx x(y+z)(1+\delta_1+\delta_2)\\&=x(y+z)(1+\delta_3) &&|\delta_3|\leqslant2^{-23}\end{aligned}$$

因为 Marc-32 是一台假想计算机，所以关于它如何计算和存储浮点数，我们可以作任何想象的假定．对一台真实的计算机，我们有关$\mathrm{fl}(x\odot y)$所作的假定非常接近事实并且能被用于可靠的误差估计．然而瞬间的反省表明在把数舍入得到机器数之前，计算机不可能完全精确地形成所有结合$x\odot y$．我们在例1中看到2/3不能以浮点形式精确计算．在实际执行中，许多计算机在专用寄存器中进行算术运算，这种寄存器要比通常的机器数有更多的位．这些额外的位称为**保护位**，它使得数以额外精度暂时存在．当然，在这个专用寄存器中对一个数应用舍入过程来产生机器数．保护位的位数和其他细节因计算机不同而不同，而且要确切地了解一台特定的计算机怎样处理这些事情有时是困难的．这个主题在 Sternbenz[1974]和计算文献资料的众多论文中有进一步的研究．也见 Feldstein and Turner[1986]、Gregory[1980]、Rall[1965]、Scott[1985]以及 Waser and Flynn[1982].

2.1.10 相对误差分析

下面我们介绍一个说明怎样分析长计算的相对舍入误差的定理．这个定理大致讲述当计算$n+1$个正机器数之和时，相对误差不超过$n\varepsilon$，其中ε是所用计算机的单位舍入误差．（因为n是所有加数的个数，所以很容易记忆这个结论．）

定理 1(加法的相对舍入误差定理) 设 x_0, x_1, …, x_n 是计算机的正机器数，ε 是计算机的单位舍入误差. 则以常规方法计算

$$\sum_{i=0}^{n} x_i$$

的相对舍入误差至多是 $(1+\varepsilon)^n-1\approx n\varepsilon$.

49

证明 设 $S_k=x_0+x_1+\cdots+x_k$，且 S_k^* 是计算机计算而不是 S_k 求和的结果. 这些量的递归公式是

$$\begin{cases} S_0 = x_0 \\ S_{k+1} = S_k + x_{k+1} \end{cases} \qquad \begin{cases} S_0^* = x_0 \\ S_{k+1}^* = \mathrm{fl}(S_k^* + x_{k+1}) \end{cases}$$

为了分析，我们定义

$$\rho_k = \frac{S_k^* - S_k}{S_k} \qquad \delta_k = \frac{S_{k+1}^* - (S_k^* + x_{k+1})}{S_k^* + x_{k+1}}$$

这样，$|\rho_k|$ 就是计算机计算得到的部分和 S_k^* 近似前 k 项部分和 S_k 的相对误差，而 $|\delta_k|$ 是量 $\mathrm{fl}(S_k^*+x_{k+1})$ 近似 $S_k^*+x_{k+1}$ 的相对误差. 利用 ρ_k 和 δ_k 定义的等式，有

$$\begin{aligned} \rho_{k+1} &= (S_{k+1}^* - S_{k+1})/S_{k+1} \\ &= [(S_k^* + x_{k+1})(1+\delta_k) - (S_k + x_{k+1})]/S_{k+1} \\ &= \{[S_k(1+\rho_k) + x_{k+1}](1+\delta_k) - (S_k + x_{k+1})\}/S_{k+1} \\ &= \delta_k + \rho_k(S_k/S_{k+1})(1+\delta_k) \end{aligned}$$

因为 $S_k<S_{k+1}$ 和 $|\delta_k|\leqslant\varepsilon$，我们能断定

$$|\rho_{k+1}| \leqslant \varepsilon + |\rho_k|(1+\varepsilon) = \varepsilon + \theta|\rho_k|$$

其中取 $\theta=1+\varepsilon$. 于是，得到相继的不等式

$$\begin{aligned} |\rho_0| &= 0 \\ |\rho_1| &\leqslant \varepsilon \\ |\rho_2| &\leqslant \varepsilon + \theta\varepsilon \\ |\rho_3| &\leqslant \varepsilon + \theta(\varepsilon + \theta\varepsilon) = \varepsilon + \theta\varepsilon + \theta^2\varepsilon \\ &\vdots \end{aligned}$$

一般结果是

$$\begin{aligned} |\rho_n| &\leqslant \varepsilon + \theta\varepsilon + \theta^2\varepsilon + \theta^3\varepsilon + \cdots + \theta^{n-1}\varepsilon \\ &= \varepsilon(1+\theta+\cdots+\theta^{n-1}) \\ &= \varepsilon[(\theta^n-1)/(\theta-1)] \\ &= \varepsilon\{[(1+\varepsilon)^n-1]/\varepsilon\} \\ &= (1+\varepsilon)^n - 1 \end{aligned}$$

用二项式定理，我们有

$$(1+\varepsilon)^n - 1 = 1 + \binom{n}{1}\varepsilon + \binom{n}{2}\varepsilon^2 + \binom{n}{3}\varepsilon^3 + \cdots - 1 \approx n\varepsilon$$ ■

50

例 3 假设数 x 由无穷级数

$$x = \sum_{k=1}^{\infty} s_k$$

来定义，这里 s_k 是给定的实数. 我们打算分两个阶段来近似 x. 首先，对于大的 n 值，计算部分和

$$x_n = \sum_{k=1}^{n} s_k$$

然后，通过保留小数点后面一定的位数舍入 x_n，比如保留 m 位. 那么能否确定末尾的位数(小数点后面第 m 个位)是正确的?

解 设 $\widetilde{x}_n$ 是 x_n 的舍入值. 我们希望

$$|\widetilde{x}_n - x| \leqslant \frac{1}{2} \times 10^{-m}$$

若 x_n 被恰当地舍入到 $\widetilde{x}_n$，则

$$|\widetilde{x}_n - x_n| \leqslant \frac{1}{2} \times 10^{-m}$$

但是

$$|\widetilde{x}_n - x| \leqslant |\widetilde{x}_n - x_n| + |x_n - x| \leqslant \frac{1}{2} \times 10^{-m} + |x_n - x|$$

这个不等式无法改进，它显示除非 $x_n = x$，否则不可能获得 $|\widetilde{x}_n - x| \leqslant 10^{-m}/2$.

比较现实的要求是 $|\widetilde{x}_n - x| \leqslant (6/10) \times 10^{-m}$. 于是我们对 x_n 要求

$$\frac{1}{2} \times 10^{-m} + |x_n - x| < \frac{6}{10} \times 10^{-m} \tag{7}$$

或

$$|x_n - x| < 10^{-m-1} \tag{8}$$

因此，我们能得到数 n. 这个不等式等价于

$$\left| \sum_{k=n+1}^{\infty} s_k \right| < 10^{-(m+1)}$$

■

习题 2.1

[51]

1. 如果 Marc-32 不把数正确地舍入而是简单地截断超过的位，试问单位舍入是多少?
2. 如果 1/10 被正确地舍入到规格化二进制数$(1.a_1a_2\cdots a_{23})2 \times 2^m$，试问舍入误差是多少? 相对误差是多少?
3. a. 如果 3/5 被正确地舍入到二进制数$(1.a_1a_2\cdots a_{24})_2$，试问相对误差是多少?
   b. 对于数 2/7 回答与 a 相同的问题.
4. $2(1-2^{-24})/3$ 是否为 Marc-32 中的机器数? 说明理由.
5. 设 $x_1, x_2, \cdots, x_n$ 是 Marc-32 中的正机器数. S_n 表示和 $x_1 + x_2 + \cdots + x_n$，而 S_n^* 表示相应的计算机计算的和.(假设按给定的次序执行加法.) 证明：若对每个 i，$x_{i+1} \geqslant 2^{-24} S_i$，则
$$|S_n^* - S_n| / S_n \leqslant (n-1)2^{-24}$$
6. 对于 Marc-32 中数的表示，证明不等式(6)的微小改进式
$$\left| \frac{x - x^*}{x} \right| \leqslant \frac{1}{1 + 2^{24}}$$

7. 在 Marc-32 中有多少个规格化机器数？(不包括 0.)

8. Marc-32 中的每个机器数是否都有唯一的规格化表达式？

9. 设 $x=(1.11\cdots111\ 000\cdots)_2\times 2^{16}$，其中小数部分首先有 26 个 1，其后都是 0. 对 Marc-32，求 x_-，x_+，$fl(x)$，$x-x_-$，x_+-x，x_+-x_-，$|x-fl(x)|/x$.

10. 设 $x=2^3+2^{-19}+2^{-22}$. 找出 Marc-32 上的机器数，使得它们恰好分别在 x 的左右边. 求 $fl(x)$，绝对误差 $|x-fl(x)|$，相对误差 $|x-fl(x)|/|x|$. 并验证这种情况的相对误差不超出 2^{-24}.

11. 在带 43 位规格化尾数的二进制计算机中，找出恰好在 1/9 右边的机器数.

12. 如果 $x=\sum_{n=1}^{26}2^{-n}$ 而 x^* 是 Marc-32 上最接近 x 的机器数，试问 x^*-x 的准确值是多少？

13. 设 $S_n=x_1+x_2+\cdots+x_n$，其中每个 x_i 都是机器数，而 S_n^* 是计算机计算的相应值. 则 $S_n^*=fl(S_{n-1}^*+x_n)$. 证明在 Marc-32 上，

$$S_n^*\approx S_n+S_2\delta_2+\cdots+S_n\delta_n\qquad|\delta_k|\leqslant 2^{-24}$$

14. 在 Marc-32 上，下列哪些式子未必成立？(这里 x，y 和 z 都是机器数并且 $|\delta|\leqslant 2^{-24}$.)
   a. $fl(xy)=xy(1+\delta)$
   b. $fl(x+y)=(x+y)(1+\delta)$
   c. $fl(xy)=xy/(1+\delta)$
   d. $|fl(xy)-xy|\leqslant|xy|2^{-24}$
   e. $fl(x+y+z)=(x+y+z)(1+\delta)$

15. 对 Marc-32 上的机器数 a，b，c，d，求计算 $(ab)/(cd)$ 时产生的相对误差的界.

16. 下列各数是否为 Marc-32 中的机器数？
   a. 10^{40}
   b. $2^{-1}+2^{-26}$
   c. $1/5$
   d. $1/3$
   e. $1/256$

52

17. 设 $x=2^{16}+2^{-8}+2^{-9}+2^{-10}$，而 x^* 是 Marc-32 中最接近 x 的机器数. 试问 $|x-x^*|$ 是多少？

18. 评论下列论断：在 Marc-32 上，当算术上结合两个机器数时，相对舍入误差不会超过 2^{-24}. 因此，当结合 n 个这样的数时，相对舍入误差不会超过 $(n-1)2^{-24}$.

19. 设 $x=2^{12}+2^{-12}$.
   a. 在 Marc-32 中找出机器数 x_- 和 x_+ 使得它们恰好分别在 x 的左右边.
   b. 对此数，证明在 Marc-32 中，x 和 $fl(x)$ 间的相对误差不大于单位舍入误差.

20. 在计算 Marc-32 中 n 个机器数之积时，相对舍入误差可能是多少？如果这 n 个数不一定都是机器数(但是在此计算机的范围之内)，你的回答有怎样变化？

21. 给出实数 x，y 使得 $fl(x\odot y)\neq fl(fl(x)\odot fl(y))$ 的一些例子. 用一台带 5 位小数的计算机说明所有 4 种算术运算.

22. 当我们写 $\prod_{i=1}^{n}(1+\delta_i)=1+\varepsilon$，$|\delta_i|\leqslant 2^{-24}$ 时，ε 的可能值范围是多少？$|\varepsilon|\leqslant n2^{-24}$ 是否为实际界？

23. 假设数 z_1，z_2，…是由数据 x，a_1，a_2，…通过算法

$$\begin{cases}z_1=a_1\\ z_n=xz_{n-1}+a_n\qquad(n\geqslant 2)\end{cases}$$

(这是霍纳算法)计算得到的. 假定这些数据是机器数. 证明用计算机产生的 z_n 是由把精确算法应用到扰动数据上产生的数. 给出关于计算机的单位舍入误差的扰动界.

24. 在本节的定理中出现一个量$(1+\varepsilon)^n-1$. 证明：若 $\varepsilon n<0.01$，则$(1+\varepsilon)^n-1<0.010\,06$.
25. 根据课本中给定的假设，建立(7)式和(8)式.
26. 在 Marc-32 中，2 的相继幂之间存在多少个浮点数?
27. 除了正整数外，还有什么数可用来作为一个数系的基数? 例如，我们能否用 π?（例如，见 Rousseau [1995].）
28. 带 48 位尾数的二进制计算机的单位舍入误差是多少?
29. 在一台指定尾数为 12 位小数的十进制计算机中，单位舍入误差是多少? 这样的计算机以 $x=\pm r\times 10^n$，$1/10\leqslant r<1$ 的形式存储数.
30. 证明在 Marc-32 中，4/5 不能被准确地表示. 其最接近的机器数是多少? 在 Marc-32 中存储这个数涉及的相对舍入误差是多少?
31. 什么数能用二进制有限表示而在十进制中不行?

53

32. n 个机器数相加的相对舍入误差是多少?（不假定这些数都是正的，因为课本中的定理包含这种情况.）
33. 在 Marc-32 的范围中找一个实数 x 使得 $\mathrm{fl}(x)=x(1+\delta)$，其中 $|\delta|$ 尽可能大. 能否由 $|\delta|$ 来获得界 2^{-24}?
34. 证明在对 Marc-32 所做的假设下，$\mathrm{fl}(x)=x/(1+\delta)$，其中 $|\delta|\leqslant 2^{-24}$.
35. 证明：若 x 是一台计算机的浮点机器数，而这台计算机有单位舍入 ε. 则 $\mathrm{fl}(x^k)=x^k(1+\delta)^{k-1}$，其中 $|\delta|\leqslant\varepsilon$.
36. 举例说明通常对机器数 x，y 和 z，有 $\mathrm{fl}[\mathrm{fl}(xy)z]\neq\mathrm{fl}[x\mathrm{fl}(yz)]$. 这种现象常常非正式地被说成是*机器乘法不满足结合律*.
37. 证明：若 x 和 y 是 Marc-32 的机器数，并且 $|y|\leqslant|x|2^{-25}$，则 $\mathrm{fl}(x+y)=x$.
38. 假设 x 是机器数且 $-\infty<x<0$. 在 IEEE 标准算术运算中，计算机计算 $-\infty+x$，$-\infty*x$，$x/-\infty$ 和 $-\infty/x$ 返回的是怎样的值?
39. 求下列循环小数的值.

    a. 0.181 818…

    b. 2.702 702 702 7…

    c. 98.198 198 198 1…
40. 固定一个整数 N，若实数 $x=q2^n$，其中 $1/2\leqslant q<1$，$|n|<N$，则称 x 为*可表示的*. 证明若 x_1，x_2，…，x_k 是可表示的，并且它们的积是可表示的，则 uv 也是可表示的，其中 $u=\max(x_i)$，$v=\min(x_i)$.
41. 若一台计算机用 β 作基数并且有 n 位尾数，考虑机器运算中适当舍入的机器数 ε，在课本中被定义为 $\varepsilon=\beta^{1-n}/2$. 证明 ε 是满足不等式 $\mathrm{fl}(1+\varepsilon)>1$ 的最小机器数.

计算机习题 2.1

1. 不用 Fortran 90 的内部过程，编写一个在单精度和双精度中都能使用的程序来计算你的计算机的机器精度值 ε. 这是一个精确值还是近似值? *提示*：求具有形式 2^{-k}的最小正机器数 ε 使得 $1.0+\varepsilon\neq 1.0$.
2. (续)对最大和最小机器数重复上题.
3. 学生有时会混淆数 tiny 和 epsilon 之间的差别. 请解释它们的差别. 然后设计并运行一个数值计算机实验来展示这一区别.
4. 反复用 2 相除并打印结果，你可能观察到似乎能获得比最小机器数 tiny 更小的实数. 请解释为什么这是不

可能的，并且说明出现了什么情况.

5. 显然，通过检查实数在计算机表示中的一个位，人们能确定此数的符号. 类似地，人们仅从一个位就能确定一个整数是偶数还是奇数. 请说明为什么这是可能的. 设计一个计算机实验来说明这种情况.
6. 令 X=1.0/3.0，打印它的内部计算机表示以及与它对应的存储于计算机中的十进制小数. 对十进制小数，使用大号格式域. 解释并讨论这些结果.

54

7. 把 97.6 和 12.9 读入并回显在屏幕上. 接着，用较大数减较小数并打印结果. 开始时，使用默认打印格式，然后用大号格式域. 请讨论这些结果.
8. a. 证明 IEEE 双精度浮点数等距地位于带间隙 $\epsilon=2^{-52}$ 的区间[1, 2]中. 换言之，在 1 和 2 之间的双精度浮点数可以表示为 $x=1+\epsilon k$，其中 $k=1, 2, \cdots, 2^{52}-1$ 并且 $\epsilon=2^{-52}$.
   b. 证明在区间[1/2, 1]中包含同样数目的间隙为 $\epsilon/2$ 的浮点数.
9. a. 找出任意 IEEE 双精度浮点数 x，$1<x<2$ 使得 $x*(1/x)\neq 1$；即 $\mathrm{fl}(x\mathrm{fl}(1/x))$恰好不是 1.
   b. 使用强力搜索找出上面这些数中的最小数.
   (Edelman[1994]显示怎样用数学分析来充分地帮助搜索.)

2.2 绝对误差和相对误差：有效位丢失

当实数 x 被另一数 x^* 近似时，**误差**是 $x-x^*$. **绝对误差**是

$$|x-x^*|$$

而**相对误差**是

$$\left|\frac{x-x^*}{x}\right|$$

在科学测量中，有意义的几乎总是相对误差. 对于一个正在被测量的量，如果不了解它的大小，那么常常很少使用有关绝对误差的信息. (在确定木星到地球的距离时，仅仅 1 米的误差是相当寻常的，但是你不会希望一个外科医生在切割手术中产生如此的误差!)

在舍入误差研究中我们已经考虑了相对误差. 不等式

$$\left|\frac{x-\mathrm{fl}(x)}{x}\right|\leqslant\epsilon$$

表述用实数 x 接近的浮点机器数表示 x 所涉及的相对误差.

2.2.1 有效位丢失

虽然舍入误差是不可避免并且难以控制的，但是由计算产生的其他类型误差则在我们的掌控之中. 数值分析的主题很大程度上贯注理解和控制各种类型的误差. 这里我们将处理一种典型的误差，其常常是由粗心的程序设计所引起的.

55

为了看看这种有较大相对误差出现的情况，我们考虑两个彼此接近的数相减. 例如，

$$x = 0.372\ 147\ 869\ 3$$
$$y = 0.372\ 023\ 057\ 2$$
$$x-y = 0.000\ 124\ 812\ 1$$

如果这个计算在有 5 位尾数的十进制计算机上被执行，我们会看到

$$\mathrm{fl}(x) = 0.372\ 15$$
$$\mathrm{fl}(y) = 0.372\ 02$$
$$\mathrm{fl}(x)-\mathrm{fl}(y) = 0.000\ 13$$

因而相对误差非常大：

$$\left|\frac{x-y-[\mathrm{fl}(x)-\mathrm{fl}(y)]}{x-y}\right|=\left|\frac{0.000\,124\,812\,1-0.000\,13}{0.000\,124\,812\,1}\right|\approx 4\%$$

只要计算机移动尾数的数位来获得一个规格化浮点数，就会在右边添0．这些0是虚假的，而且并不表示增加精度．因而，$\mathrm{fl}(x)-\mathrm{fl}(y)$在计算机中被表示为$0.130\,00\times10^{-3}$，尾数中这些0仅仅起着占位符作用

2.2.2　几乎相等量的减法

作为一个法则，我们应该避免用几乎相等量的减法，这种减法将危及精度．一个谨慎的程序员应该警觉这种情况．为了说明重新编程能起什么作用，下面来看一个例子．

例1　对小的数值x，赋值语句

$$y\leftarrow\sqrt{x^2+1}-1$$

涉及减法相消和有效位丢失．我们怎样能避免这个困难？

解　以下列方式重写函数

$$y=(\sqrt{x^2+1}-1)\left(\frac{\sqrt{x^2+1}+1}{\sqrt{x^2+1}+1}\right)=\frac{x^2}{\sqrt{x^2+1}+1}$$

于是，通过对不同的赋值语句

$$y\leftarrow x^2/(\sqrt{x^2+1}+1)$$

56　重新编程而避免了这种困难．■

2.2.3　精度丢失

一个有趣的问题是：当x接近y时，在减法$x-y$中确切地丢失了多少位有效的二进制位？准确答案要依赖于x和y的精确值．然而，我们可以根据$|1-y/x|$这个量，获得界，量$|1-y/x|$是x与y接近程度的一种方便的估计．下面的定理中包含有用的上下界．（这个定理与机器无关．）

定理1（精度丢失定理）　*若x和y是正的规格化浮点二进制机器数使得$x>y$并且*

$$2^{-q}\leqslant 1-\frac{y}{x}\leqslant 2^{-p}$$

则在减法$x-y$中丢失至多q个、至少p个有效的二进制位．

证明　我们将证明下界，把上界留作练习．x和y的规格化二进制浮点形式是

$$x=r\times2^n\qquad(\frac{1}{2}\leqslant r<1)$$

$$y=s\times2^m\qquad(\frac{1}{2}\leqslant s<1)$$

因为x大于y，所以在执行x减y之前，计算机必须对y移位以便它们有相同的指数．所以，必须把y写成

$$y=(s\times2^{m-n})\times2^n$$

于是我们有

$$x - y = (r - s \times 2^{m-n}) \times 2^n$$

这个数的尾数满足

$$r - s \times 2^{m-n} = r\left(1 - \frac{s \times 2^m}{r \times 2^n}\right) = r\left(1 - \frac{y}{x}\right) < 2^{-p}$$

为使 $x-y$ 的计算机表示规格化，至少需要向左移动 p 位. 因而至少有 p 个虚假的 0 被添加到尾数的右端，这意味着至少丢失了 p 个二进制的精度. ■

例 2 考虑赋值语句

$$y \leftarrow x - \sin x$$

因为对小的 x 值 $\sin x \approx x$，所以这个计算涉及有效位丢失. 怎样才能避免这种情况？ 57

解 让我们找出函数 $y = x - \sin x$ 的另一种形式. 这里可以利用 $\sin x$ 的泰勒级数. 因而，我们有

$$\begin{aligned} y &= x - \sin x \\ &= x - \left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots\right) \\ &= \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!} - \cdots \end{aligned}$$

若 x 接近于 0，则在这个赋值语句中可以使用截断级数

$$y \leftarrow (x^3/6)(1 - (x^2/20)(1 - (x^2/42)(1 - x^2/72)))$$

在这个函数中，如果 x 的取值范围较大，那么求 y 的值时最好把两个赋值语句都用上，每个在其适当的范围内使用. ■

在这个例子中，需要进一步分析才能确定每个赋值语句中 x 值的适当范围. 用关于精度丢失的定理 1，我们看到在第一个赋值语句的减法中通过限制 x，位的丢失可以限定在最多一位使得

$$1 - \frac{\sin x}{x} \geqslant \frac{1}{2}$$

(这里我们只考虑当 $\sin x > 0$ 时的情况.)用计算器，容易确定 x 必须至少是 1.9. 因此对 $|x| \geqslant 1.9$，我们应该用含有 $x - \sin x$ 的第一个赋值语句，而对 $|x| < 1.9$，则应该用截断级数. 可以验证对于最差的情况($x=1.9$)，级数中的 7 项产生的 y，其误差最多是 10^{-9}. (见习题 2.2.1.)

为了用前面级数中一定数量的项构造 $y = x - \sin x$ 的子程序，使用公式

$$\begin{cases} t_1 = x^3/6 \\ t_{n+1} = -t_n x^2/[(2n+2)(2n+3)] \qquad (n \geqslant 1) \end{cases}$$

其中每项都能从前项导出. 部分和

$$s_n = \sum_{k=1}^{n} t_k$$

可由

$$\begin{cases} s_1 = t_1 \\ s_{n+1} = s_n + t_{n+1} \qquad (n \geqslant 1) \end{cases}$$

归纳地得到.

许多计算都利用**双精度**来避免或改善有效位丢失. 在这种计算模式中, 每个实数被分配两个存储字. 这至少使尾数位数加倍. 计算中的某些极重要部分用双精度来执行, 而其余部分用单精度来执行. 这比整个问题都用双精度来执行要经济些, 因为后面的模式增加计算时间(从而增加成本)2至4倍. 这就是为什么双精度算术运算通常用软件执行而单精度算术运算由硬件执行的原因.

[58]

2.2.4 函数求值

还存在另一种情况: 出现极端的有效位丢失. 这种情况通常在对非常大的自变量求某些函数值的过程中出现. 我们用余弦函数来说明, 此函数具有周期性

$$\cos(x + 2n\pi) = \cos x \qquad (n \text{ 是整数})$$

利用这个性质, 通过求区间$[0, 2\pi]$内的**约化自变量**的值来实现求任意自变量的$\cos x$值. 计算机的库存子程序在一个称为**值域约化**的进程中利用了这种性质. 其他的一些性质, 比如,

$$\cos(-x) = \cos x = -\cos(\pi - x)$$

也可以使用. 例如, 通过找出约化自变量的值

$$y = 33\ 278.21 - 5\ 296 \times 2\pi = 2.46$$

着手求$\cos x$在$x = 33\ 278.21$处的值. 这里我们仅保留2位十进制小数, 因为在初始自变量中只出现精确的2位小数. 尽管初始自变量有7位有效数字, 但这个约化自变量只有3位有效数字. 那么余弦最多有3位有效数字. 我们不应当误认为$5\ 296 \times 2\pi$中的无限精度被传递给了约化自变量y. 此外, 人们也不应该被从一个子程序中打印输出的明显精度所欺骗. 若这余弦子程序给出一个具有3位有效数字的自变量y, 则值$\cos y$不会有多于3位的有效数字, 即使它可能被显示为

$$\cos(2.46) = -0.776\ 570\ 283\ 5$$

(这就是为什么子程序把自变量看作精确到全机器精度的原因, 当然它没有这样的精度.)

2.2.5 区间算术运算

根据已知的舍入误差范围来控制计算的方法是**区间算术运算**. 按这种计算方式, 每个计算的数伴有一个保证含有正确值的区间. 当然从理想上来说, 这些区间非常小, 并且最终得到的解仅有微小的误差. 然而, 在整个冗长的计算中携带区间(代替简单的机器数)进行肯定不方便. 因此, 仅在要求计算非常可靠时才使用它. 另外, 阻止区间逐渐变得大大超过现实情况可能是非常困难的. 最近, 许多有效地提供在计算中使用区间算术运算的软件包已被开发. 区间算术运算有它自己的文献, 包括供研究的期刊. 关于这方面的书有 Alefeld and Grigorieff [1980]、Alefeld and Herzberger [1983]、Kulisch and Miranker [1981] 以及 Moore [1966, 1979]. 关于区间算术运算的最新研究进展可以在因特网中的主页上找到. (见附录A, "数学软件一览".)

[59]

习题 2.2

1. 利用泰勒定理的误差项, 证明若例2中的误差不超过10^{-9}, 则其级数至少需要7项.
2. 当我们对$x = 1/2$执行减法$x - \sin x$时, 在计算机上丢失多少精确位?

3. 当 $x=1/4$ 时，在减法 $1-\cos x$ 中丢失多少精确位？
4. (续)对于上题中的函数，找一个适当的能对它精确计算的泰勒级数.
5. 找一个能调用系统函数 $\sin x$ 或 $\cos x$ 的适当的三角恒等式以便对微小的 x，能精确计算 $1-\cos x$.（有两个有效的答案.）
6. 找一种计算 $\sqrt{x^4+4}-2$ 的方法，要求没有不适当的有效位丢失.
7. 利用定义 $\sinh x \equiv (e^x - e^{-x})/2$，讨论计算 $\sinh x$ 的问题.
8. 在解二次方程 $ax^2+bx+c=0$ 过程中利用公式

$$x = (-b \pm \sqrt{b^2-4ac})/2a$$

当 $4ac$ 相对 b^2 是微小时，

$$\sqrt{b^2-4ac} \approx |b|$$

因此存在有效位丢失. 请提出一种解决这个问题的方法.

9. 提出避免在下列计算中丢失有效位的方法.
   a. $\sqrt{x^2+1}-x$
   b. $\log x - \log y$
   c. $x^{-3}(\sin x - x)$
   d. $\sqrt{x+2}-\sqrt{x}$
   e. $e^x - e$
   f. $\log x - 1$
   g. $(\cos x - e^{-x})/\sin x$
   h. $\sin x - \tan x$
   i. $\sinh x - \tanh x$
   j. $\ln(x+\sqrt{x^2+1})$（提示：这是函数 $\sinh^{-1} x$.）

60

10. 对任意 $x_0 > -1$，由

$$x_{n+1} = 2^{n+1}\left[\sqrt{1+2^{-n}x_n}-1\right]$$

递归定义的序列收敛于 $\ln(x_0+1)$.（见 Henrici[1962，第 243 页].）用一种可以避免有效位丢失的方法改写这个公式.

11. 为了有利于计算 x 接近于 0 时 $\tan x - \sin x$ 的值，请改写下列公式.
    a. $\sin x[(1/\cos x)-1]$
    b. $x^3/2$
    c. $(\sin x)/(\cos x) - \sin x$
    d. $(x^2/2)(1-x^2/12)\tan x$
    e. $x^2 \tan x/2$
    f. $\tan x \sin^2 x/(\cos x+1)$
12. 寻找不严重丢失有效位的方法来计算下列函数.
    a. $(1-x)/(1+x)-1/(3x+1)$
    b. $\sqrt{x+(1/x)}-\sqrt{x-(1/x)}$
    c. $e^x - \cos x - \sin x$
13. 由级数

$$e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \cdots$$

来讨论 e^{-x}，$x>0$ 的计算. 假定不利用系统函数 e^x，试提出一个较好的方法.

14. 对 x 的某些值在计算 $f(x)=1+\cos x$ 过程中存在减法相消的问题. 试问这些 x 的值是什么并且怎样防止精度丢失?

15. 考虑函数 $f(x)=x^{-1}(1-\cos x)$.

    a. $f(0)$的正确定义是什么，即，使 f 连续的值是什么?

    b. 若使用已知公式，则在哪些点附近存在有效位丢失?

    c. 我们怎样克服 b 中的困难? 找出一种不用泰勒级数的方法.

    d. 当你在 c 中给出的新公式在某个其他点上含有减法相消时，试描述怎样避免那个麻烦.

16. 设 $f(x)=-e^{-2x}+e^x$. 对微小的 x 值，这些关于 f 的公式 x，$3x$，$3x(1-x/2)$，$2-3x$，或 $e^x(1-e^{-3x})$中哪个最精确?

17. 若在计算 $y=\sqrt{x^2+1}-1$ 中至多丢失 2 位精度，试问必须对 x 怎样限制?

18. 为使得近似 $\sin\theta\approx\theta$ 给出的结果能精确(舍入)到 3 位十进制小数，问 θ 值范围.

19. 对于小的 x 值，近似 $\cos x\approx 1$ 究竟有多好? x 必须怎样小此近似才能有 $10^{-8}/2$ 的精度?

20. 级数 $\sum_{k=1}^{\infty} k^{-1}$ 称为**调和级数**. 它发散. 部分和 $S_n=\sum_{k=1}^{n} k^{-1}$ 能由 $S_n=S_{n-1}+n^{-1}$，$S_1=1$ 递归地计算得到. 假如在你的计算机上执行这个计算，那么所能得到的最大值 S_n 是多少?(不要在计算机上做这个实验，它太浪费了.) 见 Schechter[1984].

[61]

21. 对小的 x 值，找一种精确计算 $f(x)=x+e^x-e^{2x}$ 的方法.

22. a. 找一种方法来计算函数

$$f(x) = (e^{\tan x} - e^x)/x^3$$

    在 0 附近的精确值.

    b. 求$\lim_{x\to 0} f(x)$. 提示：见习题 1.2.4.

23. 解释为什么在利用近似

$$x - \sin x \approx (x^3/6)(1-(x^2/20)(1-x^2/42))$$

时，由减法引起的有效位丢失不严重?

24. 在计算无穷级数 $\sum_{n=1}^{\infty} x_n$ 时，假设要求答案绝对误差小于 ε. 当项加起来小于 ε 时停止是否可靠? 用级数 $\sum_{n=1}^{\infty}(0.99)^n$ 说明.

25. (续)假设项 x_n 是正负交错并且 $|x_n|$ 单调递减地收敛于 0，在这附加假设下，重复做上题. (用微积分中关于交错级数的定理.)

26. 证明：若 x 是 Marc-32 的机器数并且 $x>2^{25}\pi$，则计算出的 $\cos x$ 没有有效数字.

计算机习题 2.2

1. 对一列像 8^{-1}，8^{-2}，8^{-3}，…这样的 x 值，为计算

$$f(x) = \sqrt{x^2+1} - 1$$

$$g(x) = x^2/(\sqrt{x^2+1}+1)$$

编写且运行一个程序. 虽然 $f=g$，但是计算机产生不同的结果. 问哪个结果可靠哪个结果不可靠?

2. 编写且测试一个子程序，它能用来接受一个机器数 x 并且返回具有接近全机器精度的值 $y=x-\sin x$.

3. 用你的计算机，打印函数

$$f(x) = x^8 - 8x^7 + 28x^6 - 56x^5 + 70x^4 - 56x^3 + 28x^2 - 8x + 1$$
$$g(x) = (((((((x-8)x+28)x-56)x+70)x-56)x+28)x-8)x+1$$
$$h(x) = (x-1)^8$$

在覆盖区间[0.99, 1.01]的101个等距点处的值. 直接计算每个函数，不要重新整理或因式分解. 注意这3个函数是完全一致的. 说明打印的值不都是正的原因，尽管这些值应该是正的. 若有绘图机，则对函数值使用放大比例尺，绘制这些函数在1.0附近的图以看清有关的变化.（见 Rice[1992，第43页].）

4. 编写且测试一个提供 $1-\cos x$，$-\pi \leqslant x \leqslant \pi$ 精确值的代码. 在0附近用泰勒级数，而在其他地方，用余弦子程序. 为保证至多丢失一个二进制位，仔细确定每种方法的可用范围.

5. 编写且测试 $f(x)=x^{-2}(1-\cos x)$ 的一个函数子程序. 避免在所有自变量 x 的减法中丢失有效位并且处理在 $x=0$ 处的难点.

6. 一个有趣的数值实验是计算下面两个向量的数量积：

62

$$x = [2.718\,281\,828, -3.141\,592\,654, 1.414\,213\,562, 0.577\,215\,664\,9, 0.301\,029\,995\,7]$$
$$y = [1\,486.249\,7, 878\,366.987\,9, -22.374\,92, 4\,773\,714.647, 0.000\,185\,049]$$

以下列4种方式求和：

a. 正序 $\sum_{i=1}^{n} x_i y_i$

b. 反序 $\sum_{i=n}^{1} x_i y_i$

c. 最大-最小序(按最大到最小顺序加正数，再按最小到最大顺序加负数，然后计算两部分的和)

d. 最小-最大序(上面方法中加法的反序)

对全部8个答案既用单精度又使用双精度. 把上述结果与7位小数的正确值 $1.006\,571\times 10^{-9}$ 作比较. 并解释你的结果.

7.（续）重复上题，但是去掉 x_4 中最后的9和 x_5 中最后的7. 这个小变化会对结果有什么影响？

8. 1994年，在 Intel Pentium 计算机的有关某些大整数除法的芯片中发现了缺陷. 例如，发现

a. 5 505 001 除以 294 911 得 18.666 000 929 09

b. 4.999 999 除以 14.999 999 得 0.333 329

c. 4 195 835 除以 3 145 727 得 1.333 82

分析这些结果并且给出所涉及的绝对误差和相对误差.

9. 定义函数 $f(x, y)=9x^4-y^4+2y^2$. 我们的目标是计算 $f(40\,545, 70\,226)$. 按下面的要求执行计算：

a. 分别用整数、单精度和双精度计算.

b. 利用初等代数，证明 $f(x, y)=(3x^2-y^2+1)(3x^2+y^2-1)+1$. 并用这个公式重复 a.

c. 用函数原来的公式和增加精度的符号处理程序，首先计算精确到6位小数，然后计算精确到7，8，…，25位小数.

d. 用 b 中的公式重复 c 中的要求.

这些练习对你有何启发？

10. π 的一些有理近似值是：

$$\frac{22}{7}, \frac{333}{106}, \frac{355}{113}, \frac{104\,348}{33\,215}, \frac{1\,148\,183}{365\,478}, \frac{1\,252\,531}{398\,693}, \frac{2\,400\,714}{764\,171},$$

$$\frac{18\,057\,529}{5\,747\,890}, \frac{56\,573\,301}{18\,007\,841}, \frac{208\,235\,675}{66\,283\,474}, \frac{681\,280\,326}{216\,858\,263}$$

研究它们所涉及的绝对误差和相对误差. 用符号处理程序来计算 π 的这些有理近似值.

11. Kahan[1993]发现许多简化比如像

$$\sqrt{(e-\pi)^2}$$

这样的常量的自动化代数系统有问题. 它们有时给出 $e-\pi$，它有错误的符号，因为 $\pi>e$. 此外，他还说它们无法确定像

63

$$\sqrt{\sqrt{10}+3}\sqrt{\sqrt{5}+2}-\sqrt{\sqrt{10}-3}\sqrt{\sqrt{5}-2}-\sqrt{10\sqrt{2}+10}$$

这样的超越常量的符号. 利用符号处理程序来测试这两种情况.

12. 利用符号处理程序来找第一个大于 27 448 的素数.

2.3 稳定计算和不稳定计算：调节

本节我们介绍另一个在数值分析中反复出现的主题：**稳定**的数值过程和不稳定的数值过程之间的区别. 与之密切相关的概念是**良态**问题和**劣态**问题.

2.3.1 数值的不稳定性

通俗地讲，若一个数值过程某个阶段所产生的小误差在随后阶段中被放大从而严重降低了全部计算的精确度，则我们说这个数值过程是**不稳定**的.

一个例子有助于解释这个概念. 考虑由

$$\begin{cases} x_0 = 1 \qquad x_1 = \dfrac{1}{3} \\ x_{n+1} = \dfrac{13}{3}x_n - \dfrac{4}{3}x_{n-1} \qquad (n \geqslant 1) \end{cases} \tag{1}$$

归纳定义的实数序列. 容易看出这个递归关系产生序列

$$x_n = \left(\frac{1}{3}\right)^n \tag{2}$$

当然，(2)式对 $n=0$ 和 $n=1$ 显然成立. 假如它对 $n\leqslant m$ 成立，那么对 $n=m+1$，由

$$\begin{aligned} x_{m+1} &= \frac{13}{3}x_m - \frac{4}{3}x_{m-1} = \frac{13}{3}\left(\frac{1}{3}\right)^m - \frac{4}{3}\left(\frac{1}{3}\right)^{m-1} \\ &= \left(\frac{1}{3}\right)^{m-1}\left[\frac{13}{9}-\frac{4}{3}\right] = \left(\frac{1}{3}\right)^{m+1} \end{aligned}$$

可得它也成立. 若用这归纳定义(1)生成数值序列，比如，在 Marc-32 上，则有些计算项非常不精确. 下面是一些用类似 Marc-32 的 32 位计算机计算所得到的项：

64

$x_0 = 1.000\ 000\ 0$

$x_1 = 0.333\ 333\ 3$(正确地舍入到 7 位有效数字)

$x_2 = 0.111\ 111\ 2$(正确地舍入到 6 位有效数字)

$x_3 = 0.037\ 037\ 3$(正确地舍入到 5 位有效数字)

$x_4 = 0.012\ 346\ 6$(正确地舍入到 4 位有效数字)

$x_5 = 0.004\ 118\ 7$(正确地舍入到 3 位有效数字)

$x_6 = 0.001\ 385\ 7$(正确地舍入到 2 位有效数字)

$x_7 = 0.000\ 513\ 1$(正确地舍入到 1 位有效数字)

$$x_8 = 0.000\ 375\ 7 \text{（正确地舍入到 0 位有效数字）}$$
$$x_9 = 0.000\ 943\ 7$$
$$x_{10} = 0.003\ 588\ 7$$
$$x_{11} = 0.014\ 292\ 7$$
$$x_{12} = 0.057\ 150\ 2$$
$$x_{13} = 0.228\ 593\ 9$$
$$x_{14} = 0.914\ 373\ 5$$
$$x_{15} = 3.657\ 493 \text{（不正确，相对误差为 } 10^8 \text{）}$$

所以算法是**不稳定的**．x_n 中存在的任何误差在计算 x_{n+1} 时被乘以 13/3．因此，存在这种可能性：x_1 的误差乘上因数 $(13/3)^{14}$ 后传给 x_{15}．因为 x_1 的绝对误差约为 10^{-8}，而 $(13/3)^{14}$ 的约为 10^9，所以单独由 x_1 的误差引起的 x_{15} 的误差几乎等于 10．事实上，在计算 x_2，x_3，…中的每一个时，出现的额外舍入误差可能也都乘上各种具有形式为 $(13/3)^k$ 的因数后传给了 x_{15}．

解释这个例子的另一种方法是注意到(1)式是**差分方程**，其通解是

$$x_n = A\left(\frac{1}{3}\right)^n + B(4)^n$$

这里 A 和 B 是由初值 x_0 和 x_1 决定的常数．（读者可以参考 1.3 节线性差分方程理论）虽然我们希望计算对应于 $A=1$ 和 $B=0$ 的真解(2)，但要避免受到不需要部分 4^n 的影响是不可能的．因此，后者最终支配了想得到的解．

一个过程是数值稳定还是不稳定是以相对误差为基础来判定的．因此，如果在计算中存在大的误差，而解较大时，就完全可以接受那种情况．在前面例子中，设初值 $x_0=1$ 和 $x_1=4$．递归关系(1)不变，于是如前所述，误差仍被传递和放大．但是现在正确解是 $x_n=4^n$，并且计算结果精确到 7 位有效数字．这里列举了其中的 3 个：

$$x_1 = 4.000\ 006$$
$$x_{10} = 1.048\ 576 \times 10^6$$
$$x_{20} = 1.099\ 512 \times 10^{12}$$

65

在这种情况下，正确值大到足以掩盖误差．毫无疑问绝对误差是大的（如前所述），但是，与正确值相比较它们又相对地可忽略不计．

数

$$y_n = \int_0^1 x^n e^x \mathrm{d}x \qquad (n \geqslant 0) \tag{3}$$

的计算提供了另一个数值不稳定的例子．若我们对定义 y_{n+1} 的积分应用分部积分法，结果则得到递归关系：

$$y_{n+1} = e - (n+1)y_n \tag{4}$$

从上式和显而易见的事实 $y_0=e-1$，得到 y_1：

$$y_1 = e - y_0 = e - (e-1) = 1$$

在一台像 Marc-32 的计算机上，利用关系(4)，从 $y_1=1$ 开始，生成 y_2，y_3，…，y_{15}．这些结果中的 3 个是

$$y_2 = 0.718\ 281\ 7$$
$$y_{11} = 1.422\ 453$$
$$y_{15} = 39\ 711.43$$

这些值不可能正确. 确实，由(3)式显然可知，y 序列满足 $y_1 > y_2 > \cdots > 0$ 且 $\lim_{n\to\infty} y_n = 0$.（实际上，对 $0 < x < 1$，表达式 x^n 单调递减地趋向 0.）一旦知道这些，我们就能从(4)式看出 $\lim_{n\to\infty}(n+1)y_n = e$.

在此例中，y_2 的 δ 单位误差在计算 y_3 时被乘以 3. 3δ 的误差在计算 y_4 时被乘以 4. 由此引起的误差 12δ 在计算 y_5 时被乘以 5. 这个过程继续下去，结果在计算 y_{10} 时，误差可能差不多是 $10!\ \delta/2 \approx 2\times 10^6\delta$. 对于 y_{20}，相应的数字是 $10^{18}\delta$. 因为在 Marc-32 上 $\delta \approx 2^{-23}$，所以 $10^{18}\delta \approx 10^{10}$. 因而，这些误差完全掩盖了 y_n 的正确值；正确值迅速趋向于 0.

2.3.2 调节

条件和**调节**非正式地用于指出一个问题的解对于输入数据中的细微变化的相对敏感程度. 如果数据的微小变化能引起解的大变化，这个问题就称是**病态**的. 对于某几类问题可定义**条件数**. 若条件数较大，则问题就是病态的. 这种情况的例子稍后给出，在这里我们只讨论一些基本的例子.

虽然这里我们不以具体的方法来讨论调节，但一些基本例子还是能说明这个重要数值概念的. 已经指出，条件数与求解问题的数值解性态是密切关联的，而与特殊的解法无关. 基本上，若条件数较大，就为面临麻烦做好准备吧！

66

假设我们的问题只是简单地求函数 f 在点 x 处的值. 若 x 被略微扰动，那么对 $f(x)$ 有什么影响？当这个问题涉及绝对误差时，我们可用中值定理并且记作

$$f(x+h) - f(x) = f'(\xi)h \approx hf'(x)$$

因而，当 $f'(x)$ 不是太大时，对 $f(x)$ 扰动的影响是微小的. 然而，通常在这样的问题中重要的是**相对误差**. 在用数量 h 对 x 扰动时，我们用 h/x 作为**扰动的相对大小**. 同样地，当 $f(x)$ 被扰动到 $f(x+h)$ 时，扰动的相对大小是

$$\frac{f(x+h)-f(x)}{f(x)} \approx \frac{hf'(x)}{f(x)} = \left[\frac{xf'(x)}{f(x)}\right]\left(\frac{h}{x}\right)$$

因而，因子 $xf'(x)/f(x)$ 充当这个问题的**条件数**.

例 1 反正弦函数赋值的条件数是多少？

解 设 $f(x) = \arcsin x$，则

$$\frac{xf'(x)}{f(x)} = \frac{x}{\sqrt{1-x^2}\arcsin x}$$

对于 1 附近的 x，$\arcsin x \approx \pi/2$，并且当 x 趋向 1 时，由于这个条件数近似 $2x/(\pi\sqrt{1-x^2})$，因此它趋向无穷大. 所以，x 中微小的相对误差可能导致在 $x=1$ 附近 $\arcsin x$ 中较大的相对误差.

■

现在我们考虑函数 f 的零点(或根)定位问题.（以第 3 章中算法的观点来研究此问题.）设 f 和 g 是定义在 r 的一个邻域内的属于 C^2 类的两个函数，其中 r 是 f 的一个根.

假定 r 是**单根**，所以 $f'(r)\neq 0$. 当我们把函数 f 扰动到 $F\equiv f+\varepsilon g$ 时，问新的根在哪里？如果新的根是 $r+h$，我们将对 h 导出一个近似公式. 扰动 h 满足方程 $F(r+h)=0$ 或

$$f(r+h)+\varepsilon g(r+h)=0$$

因为 f 和 g 属于 C^2，我们可用泰勒定理表示 $F(r+h)=0$：

$$\left[f(r)+hf'(r)+\frac{1}{2}h^2f''(\xi)\right]+\varepsilon\left[g(r)+hg'(r)+\frac{1}{2}h^2g''(\eta)\right]=0$$

丢弃 h^2 项并且利用：$f(r)=0$，我们得到

$$h\approx-\varepsilon\frac{g(r)}{f'(r)+\varepsilon g'(r)}\approx-\varepsilon\frac{g(r)}{f'(r)}$$

67

例 2 我们考虑用 Wilkinson 给的经典例子来说明这种分析. 设

$$f(x)=\prod_{k=1}^{20}(x-k)=(x-1)(x-2)\cdots(x-20)$$

$$g(x)=x^{20}$$

显然 f 的根是整数 1，2，…，20. 当 f 扰动到 $f+\varepsilon g$ 时，问根 $r=20$ 受到怎样的影响？

解 答案是

$$h\approx-\varepsilon\frac{g(20)}{f'(20)}=-\varepsilon\frac{20^{20}}{19!}\approx-\varepsilon 10^9$$

因而，$f(x)$ 中 x^{20} 的系数变化 ε 可能引起根 20 总计的扰动 $10^9\varepsilon$. 所以说，这个多项式的根对系数的扰动是非常敏感的.（见计算机习题 2.3.6.） ■

还有另一类条件数与解线性方程组 $Ax=b$ 有关，这将在 4.4 节中详细讨论. 简短地说，矩阵 A 的**条件数**记作 $\kappa(A)$ 并且被定义为 A 与其逆的大小的积，即，

$$\kappa(A)=\|A\|\cdot\|A^{-1}\|$$

这里 $\|\cdot\|$ 是一个矩阵范数. 若 $Ax=b$ 的解对右边 b 的微小变化非常不敏感，则 b 中的小扰动在所计算的解 x 中仅产生小的扰动. 此时，称 A 是**良态的**. 这种情形对应于具有适当大小的条件数 $\kappa(A)$. 另一方面，若条件数较大，则 A 是**病态的**并且必然充满疑虑地接受 $Ax=b$ 的任何数值解.

n 阶**希尔伯特矩阵** $H_n=(h_{ij})$ 由 $h_{ij}=1/(i+j-1)$ 来定义，这里 $1\leqslant i\leqslant n$，$1\leqslant j\leqslant n$. 希尔伯特矩阵是一个被认真研究多年的课题. 有关它的一些历史可以看 Hestenes and Todd[1991]：《Mathematicians Learning to Use Computers》中的第 9 章. 这个矩阵因它的病态性质常常用来作测试矩阵. 事实上，$\kappa(H_n)$ 随 n 增大而迅速增大，这可从公式 $\kappa(H_n)=ce^{3.5n}$ 中看出. 这个条件数是用范数 $\|A\|=\max_{ij}|a_{ij}|$ 来计算的. 一个有意思的测试问题是对不断增加的 n 值，在一台计算机上解线性系统

$$H_nx=b$$

其中 $b_i=\sum_{j=1}^{n}h_{ij}$. 这个解应该是 $x=(1,1,\cdots,1)^T$. 对小的 n 值，计算所得的解相当精确，但是随着 n 增大，精确度迅速下降. 事实上，当 n 近似等于计算机所带的小数位数时，该数值解可能不含有效数字. 这些内容将在第 4 章中详细讨论.

68

当我们试图求表达式

$$\int_0^1 \Big[\sum_{j=0}^n a_j x^j - f(x) \Big]^2 \mathrm{d}x$$

的最小值时，在最小二乘逼近中，出现希尔伯特矩阵. 对这个表达式关于 a_i 求导并且令这些导数等于 0，我们得到**正规方程**

$$\sum_{j=0}^n a_j \int_0^1 x^i x^j \mathrm{d}x = \int_0^1 x^i f(x) \mathrm{d}x \qquad (0 \leqslant i \leqslant n)$$

因为左边的积分是

$$\frac{x^{i+j+1}}{i+j+1}\Big|_0^1 = \frac{1}{i+j+1}$$

所以正规方程的系数矩阵是 $n+1$ 阶希尔伯特矩阵. 函数 $x \to x^i$ 形成非常劣态的 n 次多项式空间的基. 对此，多项式的正交集能提供一个良好的基. 这将在 6.8 节中讨论.

习题 2.3

1. 存在一个具有下列形式的函数

$$f(x) = \alpha x^{12} + \beta x^{13}$$

其中 $f(0.1)=6.06\times10^{-13}$，$f(0.9)=0.035\,77$. 求 α 和 β，并且估计这些参数对 f 在两个指定点上微小变化的敏感性.

2. 解析地求出具有给定初值的差分方程

$$\begin{cases} x_0 = 1 \qquad x_1 = 0.9 \\ x_{n+1} = -0.2x_n + 0.99x_{n-1} \end{cases}$$

的解. 如果不是递归地计算解，是否能预测这样的计算是稳定的.

3. 指数积分是由

$$E_n(x) = \int_1^\infty (\mathrm{e}^{xt} t^n)^{-1} \mathrm{d}t \qquad (n \geqslant 0, x > 0)$$

定义的函数 E_n. 这些函数满足等式

$$nE_{n+1}(x) = \mathrm{e}^{-x} - xE_n(x)$$

若 $E_1(x)$已知，这个等式是否能用来精确地计算 $E_2(x)$，$E_3(x)$，…？提示：确定 $E_n(x)$作为 n 的一个函数，是增的还是减的.

69

4. 函数 $f(x)=x^\alpha$ 的条件数是与 x 无关的. 这个条件数是什么？

5. 下列函数的条件数是什么？在何处它们较大？

a. $(x-1)^\alpha$

b. $\ln x$

c. $\sin x$

d. e^x

e. $x^{-1}\mathrm{e}^x$

f. $\cos^{-1} x$

6. 考虑课本中的例子：$y_{n+1}=\mathrm{e}-(n+1)y_n$. 若 y_{20}要求精确到 5 位小数，试问计算 y_1，y_2，…，y_{20}需要精确到多少位小数.

7. 证明递归关系

$$x_n = 2x_{n-1} + x_{n-2}$$

具有下列形式的通解

$$x_n = A\lambda^n + B\mu^n$$

试问这个递归关系是一个从任意初值 x_0 和 x_1 出发计算 x_n 的好方法吗？

8. 斐波那契序列是由下列公式生成的

$$\begin{cases} r_0 = 1 \qquad r_1 = 1 \\ r_{n+1} = r_n + r_{n-1} \end{cases}$$

所以这个序列是 1，1，2，3，5，8，13，21，34，…. 证明序列$[2r_n/r_{n-1}]$收敛于 $1+\sqrt{5}$. 这个收敛是线性的，超线性的，还是二次的？

9. (续)当上题中的递归关系使用初值 $r_0=1$ 和 $r_1=(1-\sqrt{5})/2$ 时，试问 $r_n(n\geqslant 2)$理论上的正确值是多少？在此情况下，这递归关系能否提供一个计算 r_n 的稳定方法？

计算机习题 2.3

1. 设序列$[A_n]$和$[B_n]$由下列式子生成

$$\begin{cases} A_0 = 0 \qquad A_1 = 1 \\ A_n = nA_{n-1} + A_{n-2} \end{cases} \qquad \begin{cases} B_0 = 1 \qquad B_1 = 1 \\ B_n = nB_{n-1} + B_{n-2} \end{cases}$$

问$\lim\limits_{n\to\infty}(A_n/B_n)$是什么？

2. 贝塞尔函数 Y_n 与函数 J_n 满足同样的递归公式.（见 1.3 节.）然而，它们使用不同的初值. 对 $x=1$，它们是

$$Y_0(1) = 0.088\,256\,964\,2 \qquad Y_1(1) = -0.781\,212\,821\,3$$

用该递归公式，计算 $Y_2(1)$，$Y_3(1)$，…，$Y_{20}(1)$. 试判定这些结果是否可靠. 提示：数 $|Y_n(1)|$ 应该迅速增长. 或许你能证明像 $|Y_n(1)/Y_{n-1}(1)|>n$ 这样的不等式.

70

3. 定义

$$x_n = \int_0^1 t^n(t+5)^{-1}\,dt$$

证明 $x_0=\ln 1.2$ 并且当 $n\geqslant 1$ 时，$x_n=n^{-1}-5x_{n-1}$. 用这个递归公式计算 x_1，x_2，…，x_{10} 并且估计 x_{10} 的精度.

4. (续)寻找一种精确计算 x_{20} 的方法，或许可用截断泰勒级数来代替被积函数. 在计算 x_{20} 达到全机器精度后，用向后递归得到 x_{19}，x_{18}，…，x_0. x_0 是否正确？其他的 x_n 怎么样？当使用向后递归时，这递归关系反常吗？若有，则为什么？

5. 贝塞尔函数 $J_1(x)$，$J_2(x)$，…能用 1.3 节中的递归关系计算，如果它以 n 递减的方式来使用递归关系. 那么，我们从 $n=N$ 开始并且用公式

$$J_{n-1}(x) = \frac{2n}{x}J_n(x) - J_{n+1}(x) \qquad (N \leqslant n \leqslant 1)$$

向下进行. 为了开始，我们令 $J_{N+1}(x)=0$ 和 $J_N(x)=1$. 这些是试验值. 在计算 $J_{N-1}(x)$，$J_{N-2}(x)$，…，$J_0(x)$后，可以利用恒等式

$$J_0(x)^2 + 2\sum_{n=1}^{\infty} J_n(x)^2 = 1$$

来调节它们(即用 $\lambda J_n(x)$代替 $J_n(x)$). 显然 N 必须大到足以使对所期望的精度 $J_{N+1}(x)=0$ 是合理的. 这个方法归功于 J. C. P. Miller. 在上述过程中，取 $N=51$，计算 $J_0(1)$，$J_1(1)$，…，$J_{50}(1)$.

6. (**Perfidious 多项式**，Wilkinson[1984])

a. 利用计算机中心程序库的求根子程序，计算多项式 P 的 20 个零点，这里

$$
\begin{aligned}
P(x) = {} & x^{20} - 210x^{19} + 20\,615x^{18} - 1\,256\,850x^{17} + 53\,327\,946x^{16} \\
& - 1\,672\,280\,820x^{15} + 40\,171\,771\,630x^{14} - 756\,111\,184\,500x^{13} \\
& + 11\,310\,276\,995\,381x^{12} - 135\,585\,182\,899\,530x^{11} \\
& + 1\,307\,535\,010\,540\,395x^{10} - 10\,142\,299\,865\,511\,450x^{9} \\
& + 63\,030\,812\,099\,294\,896x^{8} - 311\,333\,643\,161\,390\,640x^{7} \\
& + 1\,206\,647\,803\,780\,373\,360x^{6} - 3\,599\,979\,517\,947\,607\,200x^{5} \\
& + 8\,037\,811\,822\,645\,051\,776x^{4} - 12\,870\,931\,245\,150\,988\,800x^{3} \\
& + 13\,803\,759\,753\,640\,704\,000x^{2} - 8\,752\,948\,036\,761\,600\,000x \\
& + 2\,432\,902\,008\,176\,640\,000
\end{aligned}
$$

使用一个能计算双精度复根的程序. 这里给出的 P 的公式是课本中所讨论的 **Wilkinson** 多项式的展开式：

$$
\begin{aligned}
p(x) = {} & (x-20)(x-19)(x-18)(x-17)(x-16)(x-15)(x-14) \\
& \times (x-13)(x-12)(x-11)(x-10)(x-9)(x-8)(x-7) \\
& \times (x-6)(x-5)(x-4)(x-3)(x-2)(x-1)
\end{aligned}
$$

71

通过计算 $|P(z_k)|$, $|p(z_k)|$, $|z_k-k|$ 来检验计算的根 z_k, $1\leqslant k\leqslant 20$. 解释原因.

b. 把 $P(x)$的 x^{20} 的系数改变为 $1+\varepsilon$，这里 $\varepsilon=10^{-2k}$，$k=8, 7, \cdots, 1$. 然后重复 a 8 次. 你得到的计算机结果与课本中给出的分析是否一致？解释原因.

c. Wilkinson 指出通过把系数 -210 改变为 $-210-2^{-23}$，根 16 和 17 变成了一对复数 $16.73\cdots\pm(2.812\cdots)i$. 重复这个数值实验.

注：20 世纪 40 年代末期，在英国的国家物理实验室，年轻的科学家 Jim Wilkinson 和 Alan Turing 等人在一台名为 Pilot ACE 计算机的新机器上合作工作. 作为这台机器的常规测试，他编写了一个使用牛顿方法计算这个多项式根的程序. 由于预测不会有问题，他用 $x_0=21$ 开始牛顿迭代并且希望立刻收敛到最大的零点 20. 当数值结果不符合这种情形时，他进一步作了研究. 最终，他被这个以及其他的数值实验引向了讨论数值数学的一个全新领域——**向后误差分析**. 有关 James Hardy Wilkinson 工作的更多细节见 L. Fox [1987]：《Biographical Memoirs of Fellows of the Royal Society》.

72

第3章　非线性方程的解

3.0　概述

本章专门讨论有关求方程**根**(或函数**零点**)的问题.

在科学工作中经常出现这类问题. 在本章中，我们关注求**解非线性**方程或非线性方程组——求 x 使得 $f(x)=0$ 或求 $X=(x_1, x_2, \cdots, x_n)^T$ 使得 $F(X)=0$. 在这些方程中，至少有一个变量以任意的非线性方式出现. 与此相反，我们将在第 4 章中讨论求形如 $Ax=b$ 的线性方程组的解.

在单实变量的实值函数这种最简单的情况下，提出的一般问题是：已知函数 f: $\mathbb{R}\rightarrow\mathbb{R}$，求 x 的值使得 $f(x)=0$. 我们将讨论一些解这种问题的标准过程.

在许多应用中能发现非线性方程的例子. 在光的衍射理论中，我们需要方程

$$x-\tan x=0$$

的根. 在行星轨道的计算中，我们需要**开普勒方程**

$$x-a\sin x=b$$

的根，其中 a 和 b 取任意值.

因为求函数零点作为研究领域的一个热点已有几百年了，所以已经建立了许多种方法. 在本章中，首先介绍三种非常有用的简单方法：对分法、牛顿法和割线法. 然后研究不动点方法和延拓法的一般理论. 此外，还讨论求多项式零点的特殊方法.

当用计算机求函数的**近似零点**时，即使精确解是唯一的，还是会出现许多近似解. 为了准确地说明这种情况，考虑多项式

$$p_4(x)=x^4-4x^3+6x^2-4x+1$$

若我们碰巧发现这多项式能被因式分解成 $p_4(x)=(x-1)^4$，则很明显 1 是唯一的零点(4 重). 假设我们得到的是展开多项式且没有看出它有这种零点. 若在一台计算机上，比如在 Marc-32 上，编写一个单精度程序并且求展开多项式在区间[0.975, 1.035]上间距为 0.001 的点的值，则会发现许多指示表示零点存在的符号变化. 用直线把这些点连起来，就得到如图 3-1 所示的

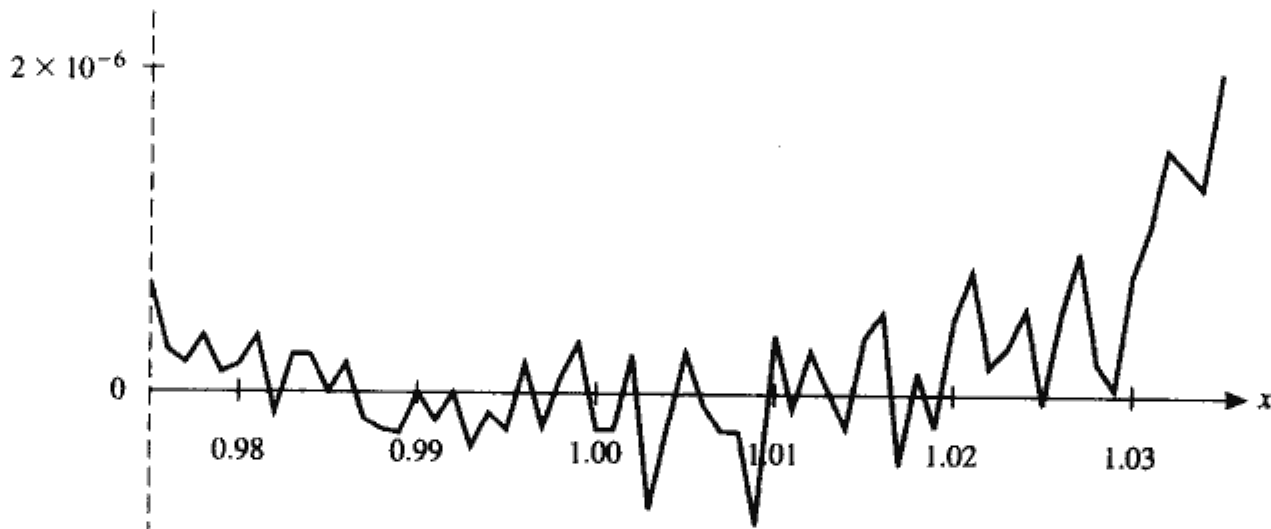


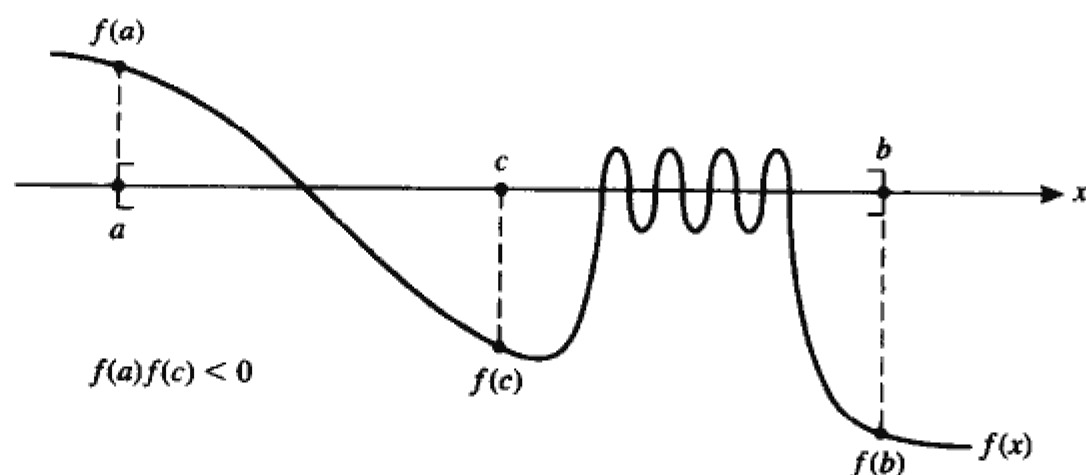
图 3-1　展开多项式的图示

图．代替多项式的一条光洁曲线，我们得到一条模糊的曲线．在区间[0.981，1.026]中的任何值都能作为真解的一个近似．其原因与多项式求值的方式、所用计算机运算的有限精度以及相关的舍入误差有关．这个例子仿效 Conte and de Boor[1980，第 73 页]中类似的例子，它说明了涉及求根中的一个不安全因素．

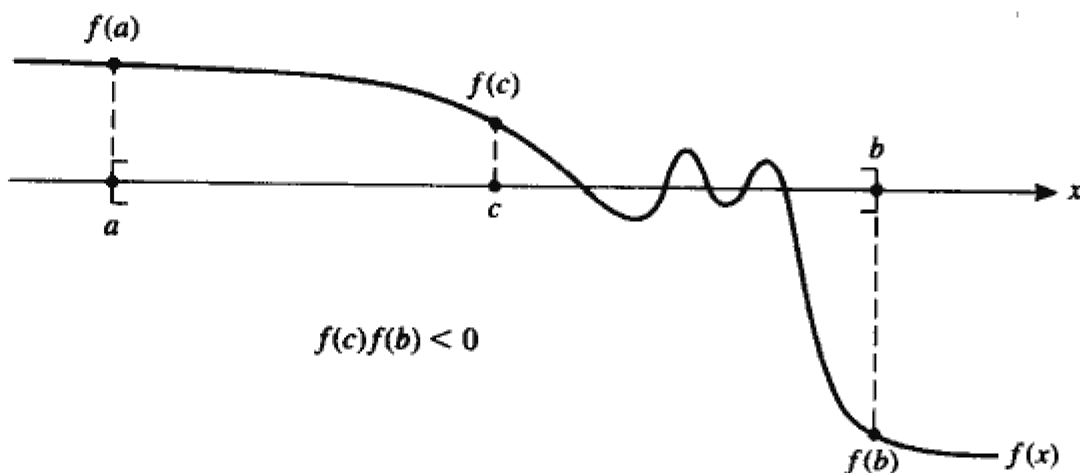
3.1 对分(区间减半)法

若 f 是区间$[a, b]$上的连续函数，且 $f(a)f(b)<0$，则 f 在(a, b)内必有一个零点．因为 $f(a)f(b)<0$，所以函数 f 在区间$[a, b]$上改变符号，因此它在这个区间内至少存在一个零点．这是中值定理(见 1.1 节)的结论．

对分法以如下的方式利用这一个思想：若 $f(a)f(b)<0$，则计算 $c=(a+b)/2$ 并且检验是否 $f(a)f(c)<0$．若这是真的，则 f 在$[a, c]$内有零点．因而把 c 改名为 b 并且在原区间一半大的新区间$[a, b]$中重新开始．若 $f(a)f(c)>0$，则 $f(c)f(b)<0$，在这种情况下我们把 c 改名为 a．不论发生哪种情况，都产生了包含 f 的零点的一个新区间，而且能重复这个过程．图 3-2a和 b 显示了这两种情况，其中假定 $f(a)>0>f(b)$．这两幅图说明对分法在区间$[a, b]$内能找到一个零点而不是全部零点的原因．当然，若 $f(a)f(c)=0$，则 $f(c)=0$ 从而求出一个零点．然而由于舍入误差，在计算机中 $f(c)$精确为 0 是完全不可能的．因此，**停止准则**不应该是 $f(c)=0$ 是否成立．必须提供一个合理的允许误差，比如在 Marc-32 上 $|f(c)|<10^{-5}$．(对假想的 Marc-32 计算机的描述见 2.1 节．) 对分法也称为**区间减半法**．



a) 对分法选择左边的子区间



b) 对分法选择右边的子区间

图 3-2

74~75

例 1 用对分法求方程 $e^x = \sin x$ 最靠近 0 的根.

解 若画出 e^x 和 $\sin x$ 的略图，可明显看出 $f(x) = e^x - \sin x$ 没有正根，并且 0 左边的第一个根在区间[−4，−3]内. 当在一台类似于 Marc-32 的计算机上从区间[−4，−3]出发执行对分算法时，产生下列数据：

k	c	$f(c)$
1	−3.500 0	−0.321
2	−3.250 0	-0.694×10^{-1}
3	−3.125 0	0.605×10^{-1}
4	−3.187 5	0.625×10^{-1}
⋮	⋮	⋮
13	−3.182 9	0.122×10^{-3}
14	−3.183 0	0.193×10^{-4}
15	−3.183 1	-0.124×10^{-4}
16	−3.183 1	0.345×10^{-5}

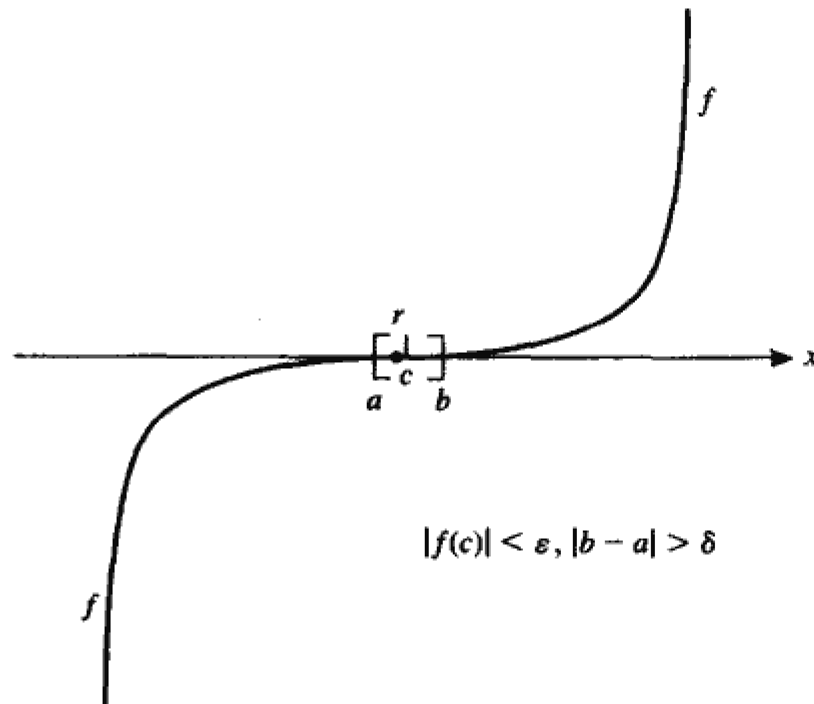
■

3.1.1 对分算法

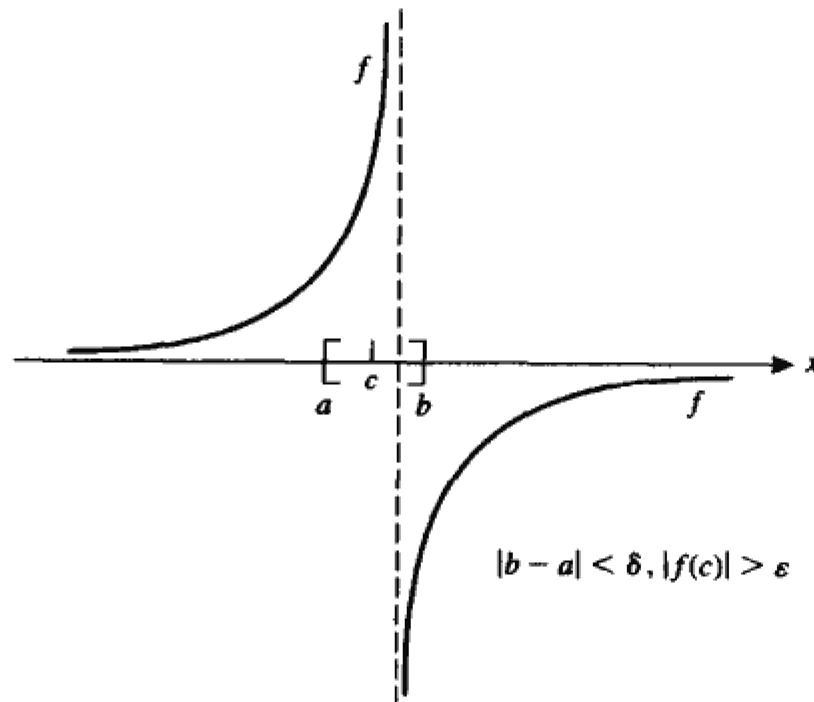
对分法的伪代码(在下面列出)中的某些部分需要加以说明. 首先，以 $c \leftarrow a + (b-a)/2$ 方式而不是以 $c \leftarrow (a+b)/2$ 方式计算中点 c. 这是为了坚持数值计算中通用策略：通过把一个小的修正项加到先前的近似值法来计算一个量是最佳的. Forsythe，Malcolm，and Moler[1977，第 162 页] 给了一个例子，这个例子表明，在一台具有有限精度的计算机上，以$(a+b)/2$ 方式计算得到的中点在区间$[a, b]$之外！其次，最好用 $\text{sign}(w) \neq \text{sign}(u)$而不是用 $wu < 0$ 来确定函数在区间上是否改变符号，因为后者需要不必要的乘法并且可能引起下溢或上溢. 再者，e 是在下面的定理中确定的计算的误差界. 最后，注意在算法中有 3 个停止准则. 首先，M 给出用户准许的最大步数. 这样的安全措施总是能降低计算进入无限循环的可能性. 其次，当误差足够小或 $f(c)$的值足够小时，停止计算. 这是由参数 δ 和 ε 控制的. 很容易给出后两个停止准则中一个满足而另一个不满足的例子. 例如，考察图 3-3a 和 b 中的两个略图. 在图 3-3a 中，零点附近的图形是平坦的，与之相应的是重根，并且对分法可能很难求出高精度的零点. 当然，图 3-3b 中的曲线不连续，然而函数的连续性可能很难预先查证. 虽然这些例子不够充分，但是我们关心的是稳健的代码，所以只要满足这 3 个停止准则中的任意一个，就停止这个算法.

下面是对分法的算法：

```
input a, b, M, δ, ε
u ← f(a)
v ← f(b)
e ← b − a
output a, b, u, v
if sign(u) = sign(v) then stop
for k = 1 to M do
    e ← e/2
    c ← a + e
```



a）准则 $|b-a|<\delta$ 失败



b）准则 $|f(c)|<\varepsilon$ 失败

图 3-3

```
        w←f(c)
        output k, c, w, e
        if |e|<δ or |w|<ε then stop
        if sign(w)≠sign(u) then
            b←c
            v←w
        else
            a←c
            u←w
        end if
end do
```

3.1.2 误差分析

为了分析对分法，用$[a_0, b_0]$，$[a_1, b_1]$等表示相继出现在对分法过程中的区间．下面是对这些数的一些观察结果：

$$\begin{aligned} &a_0 \leqslant a_1 \leqslant a_2 \leqslant \cdots \leqslant b_0 \\ &b_0 \geqslant b_1 \geqslant b_2 \geqslant \cdots \geqslant a_0 \\ &b_{n+1} - a_{n+1} = \frac{1}{2}(b_n - a_n) \qquad (n \geqslant 0) \end{aligned} \tag{1}$$

因为序列$[a_n]$是非递减的并且有上界，所以它收敛．同样，$[b_n]$也收敛．若反复应用（1）式，我们则发现

$$b_n - a_n = 2^{-n}(b_0 - a_0)$$

因而，

$$\lim_{n\to\infty} b_n - \lim_{n\to\infty} a_n = \lim_{n\to\infty} 2^{-n}(b_0 - a_0) = 0$$

若我们令

$$r = \lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n$$

则通过对不等式 $0 \geqslant f(a_n)f(b_n)$取极限，我们得到 $0 \geqslant [f(r)]^2$，由此 $f(r)=0$.

假设在这过程的某个阶段，区间$[a_n, b_n]$已被确定．若这时过程停止，则根一定会在这区间内．此时，根的最佳估计既不是 a_n 也不是 b_n 而是这个区间的中点：

$$c_n = (a_n + b_n)/2$$

于是误差有如下的界：

$$|r - c_n| \leqslant \frac{1}{2}(b_n - a_n) = 2^{-(n+1)}(b_0 - a_0)$$

76 ~ 78

总结上述讨论，我们有下面关于对分法的定理．

定理 1（对分法定理） *若$[a_0, b_0]$，$[a_1, b_1]$，…，$[a_n, b_n]$，…表示对分法中的区间，则极限$\lim_{n\to\infty} a_n$和$\lim_{n\to\infty} b_n$存在且相等，并且这个极限是 f 的一个零点．若 $r=\lim_{n\to\infty} c_n$ 且 $c_n=(a_n+b_n)/2$，则*

$$|r-c_n| \leqslant 2^{-(n+1)}(b_0-a_0) \tag{2}$$

例 2 假设对分法从区间[50，63]开始．试问需要多少步才能求出具有相对精度不超过10^{-12}的根？

解 所要求的相对精度是指

$$|r - c_n| / |r| \leqslant 10^{-12}$$

我们知道 $r \geqslant 50$，从而足以保证不等式

$$|r - c_n| / 50 \leqslant 10^{-12}$$

成立．根据定理 1，我们推断下列条件是充分的：

$$2^{-(n+1)} \times (13/50) \leqslant 10^{-12}$$

解这个关于 n 的不等式，得到 $n \geqslant 37$. ■

习题 3.1

1. 用计算器求

$$x^2 - 4x\sin x + (2\sin x)^2 = 0$$

的一个正根，并且精确到 2 位有效数字.

2. 考虑从区间[1.5，3.5]开始的对分法.

   a. 试问在这个方法的第 n 步时区间宽度是多少？

   b. 根 r 与这个区间中点之间的最大距离可能是多少？

3. 若在(单精度)Marc-32 上使用从区间[128，129]开始的对分法，试问我们能否求出具有绝对精度 $<10^{-6}$ 的根？

4. 为了确保 $|r-c_n| \leqslant \varepsilon$ 成立，求对分法中需取步数的、包含 b_0-a_0 和 ε 的公式

$$n \geqslant \frac{\log(b_0 - a_0) - \log\varepsilon}{\log 2} - 1$$

5. 为了确保所求得的根具有相对精度 $\leqslant\varepsilon$，求对分法算法中应取步数的、包含 a_0，b_0 和 ε 的公式

$$n \geqslant \frac{\log(b_0 - a_0) - \log\varepsilon - \log a_0}{\log 2} - 1$$

79 这里假定 $a_0>0$.

6. (续)若 $a_0<0<b_0$，试问上题中会出现什么情况？

7. 如果从区间[2，3]开始使用对分法，为了使求得的根具有绝对精度 $<10^{-6}$，试问必须取多少步？对于相对精度回答同样的问题. 在每种情况下，对于 Marc-32 上的单精度又怎样？

8. 设 $c_n=(a_n+b_n)/2$，$r=\lim\limits_{n\to\infty}c_n$，$e_n=r-c_n$. 其中 $[a_n, b_n]$，$n\geqslant 0$，表示对一个连续函数 f 应用对分法时相继出现的区间.

   a. 证明 $|e_n| \leqslant 2^{-n}(b_1-a_1)$.

   b. 证明当 $n\to\infty$ 时，$e_n=\mathcal{O}(2^{-n})$.

   c. $|e_0| \geqslant |e_1| \geqslant \cdots$ 是否成立？解释原因.

   d. 证明 $|c_n-c_{n+1}|=2^{-n-2}(b_0-a_0)$.

   e. 证明对所有 n 和 m，$a_m\leqslant b_n$.

   f. 证明 r 是 $\bigcap\limits_{n=0}^{\infty}[a_n, b_n]$ 中唯一的元素.

   g. 证明对所有的 n，$[a_n, b_n]\supset[a_{n+1}, b_{n+1}]$.

9. 在对分法中，区间 $[a_{n-1}, b_{n-1}]$ 被分成相等的两半，其中之一被选作下一个区间. 若 $[a_n, b_n]$ 是区间 $[a_{n-1}, b_{n-1}]$ 的左半个，则定义 $d_n=0$，否则令 $d_n=1$. 用序列 d_1，d_2，…表示由算法求得的根. 提示：首先考虑 $[a_0, b_0]=[0, 1]$ 的情况，然后考虑根的二进制表示.

10. 用前两个习题的记号，建立把 a_n，b_n，c_n，d_n 联系起来的公式.

11. 举一个满足条件 $a_0<a_1<a_2<\cdots$ 的例子(或证明不存在).

12. 举一个满足条件 $a_0=a_1<a_2=a_3<a_4=a_5<a_6=\cdots$ 的例子.

13. 在对分法中，$\lim\limits_{n\to\infty}|r-c_{n+1}|/|r-c_n|$ 是否存在？解释理由.

14. 把对分法应用到一个连续函数上，结果形成区间 $[a_0, b_0]$，$[a_1, b_1]$，等等. 设 $r=\lim\limits_{n\to\infty}a_n$. 那么下列命题中哪些是假的？

    a. $a_0\leqslant a_1\leqslant a_2\leqslant\cdots$

    b. $|r-2^{-1}(a_n+b_n)| \leqslant 2^{-n}(b_0-a_0)$ $(n\geqslant 0)$

    c. $|r-2^{-1}(a_{n+1}+b_{n+1})| \leqslant |r-2^{-1}(a_n+b_n)|$ $(n\geqslant 0)$

d. $[a_{n+1}, b_{n+1}] \subseteq [a_n, b_n]$ $(n \geq 0)$

e. 当 $n \to \infty$ 时，$|r - a_n| = \mathcal{O}(2^{-n})$

f. $|r - c_n| < |r - c_{n-1}|$ $(n \geq 1)$

15. 证明用对分法求出的点 c 是经过$(a, \text{sign}(f(a)))$和$(b, \text{sign}(f(b)))$的直线与 x 轴的交点.
16. 假设对所有 n，$|a_n - b_n| \leqslant \lambda_n |a_{n-1} - b_{n-1}|$，且 $\lambda_n < 1$. 利用 $|a_0 - b_0|$ 和 $\lambda = \max_{1 \leqslant i \leqslant n}\{\lambda_i\}$ 求 $|a_n - b_n|$ 的一个上界.

计算机习题 3.1

1. 编写且测试一个执行对分算法的子程序或过程. 对下列函数和区间测试程序.

   a. $x^{-1} - \tan x$ 在区间$[0, \pi/2]$上

   b. $x^{-1} - 2^x$ 在区间$[0, 1]$上

   c. $2^{-x} + e^x + 2\cos x - 6$ 在区间$[1, 3]$上

   d. $(x^3 + 4x^2 + 3x + 5)/(2x^3 - 9x^2 + 18x - 2)$ 在区间$[0, 4]$上

[80]

2. 求数 a 和 b(其中 $a < b$)使得数学上等价的两个计算 $c \leftarrow (a+b)/2$ 和 $c \leftarrow a + 0.5(b-a)$在你的计算机上产生不同的结果. 不要选择存在下溢或上溢的例子.
3. 求函数 $f(x) = x - \tan x$ 在区间$[1, 2]$内的根.
4. 求

$$x^8 - 36x^7 + 546x^6 - 4\,536x^5 + 22\,449x^4 - 67\,284x^3 + 118\,124x^2 - 109\,584x + 40\,320 = 0$$

   在区间$[5.5, 6.5]$内的根. 把-36改成-36.001，再重做.

5. 编写且测试一种递归形式的对分算法.

3.2 牛顿法

牛顿法是一种能在许多不同情况下应用的通用过程. 特别地，当用它来求实变量实值函数零点时，常常被称为**牛顿-拉弗森迭代**. 通常，牛顿法比对分法和割线法要快，这是因为它的收敛是二次的而不是线性或超线性的. 一旦二次收敛变得有效时，即牛顿法序列的值充分地接近根时，其收敛是如此之快以致于仅仅再需要几个数值即可. 不幸的是，这种方法并不能保证总是收敛. 所以牛顿法经常与其他较慢的方法结合形成一种数值上整体收敛的混合方法.

如 3.1 节那样，我们有一个函数 f，其零点由数值计算得出. 设 r 是 f 的零点而 x 是 r 的一个近似. 若 f''存在并且连续，则由泰勒定理，得

$$0 = f(r) = f(x+h) = f(x) + hf'(x) + \mathcal{O}(h^2)$$

其中 $h = r - x$. 若 h 较小(即 x 在 r 附近)，则有理由略去 $\mathcal{O}(h^2)$项并且在余下的方程中求 h. 若我们这样做，其结果是 $h = -f(x)/f'(x)$. 若 x 是 r 的一个近似，则 $x - f(x)/f'(x)$应该是 r 的一个更好的近似. 牛顿法从 r 的一个估计 x_0 开始，然后归纳地定义

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \qquad (n \geqslant 0) \tag{1}$$

3.2.1 牛顿算法

由 x 的一个初值开始，使用 M 步牛顿法的简单算法是

[81]

```
input x, M
y ← f(x)
```

```
output 0, x, y
for k=1 to M do
    x←x−y/f′(x)
    y←f(x)
    output k, x, y
end do
```

更详细的包含停止准则的伪代码如下：

```
input x0, M, δ, ε
v←f(x0)
output 0, x0, v
if |v|<ε then stop
for k=1 to M do
    x1←x0−v/f′(x0)
    v←f(x1)
    output k, x1, v
    if |x1−x0|<δ or |v|<ε then stop
    x0←x1
end do
```

基于这两个算法中任意一个的计算机程序需要 $f(x)$和 $f'(x)$的子程序或过程.

例 1 以双精度方式计算，用牛顿法来求函数 $f(x)=e^x-1.5-\tan^{-1}x$ 的负零点.

解 上述算法在一台尾数有 48 位的单精度计算机上以双精度执行.（双精度机器数有 96 位，对应于约 28 位十进制小数.）函数 $f'(x)=e^x-(1+x^2)^{-1}$与 f 一样，必须已被编程. 选择 $x_0=-7$ 为初始点. 来自于计算机程序的输出数据是

k	x	$f(x)$
0	−7.000 000 000 000 000 000 000 000 00	-0.702×10^{-1}
1	−10.677 096 176 640 013 992 969 843 86	-0.226×10^{-1}
2	−13.279 167 375 632 712 908 597 863 19	-0.437×10^{-2}
3	−14.053 655 854 269 238 734 748 317 53	-0.239×10^{-3}
4	−14.101 109 956 866 413 476 163 127 06	-0.800×10^{-6}
5	−14.101 269 770 939 415 946 215 795 06	-0.901×10^{-11}
6	−14.101 269 772 739 968 425 083 003 14	-0.114×10^{-20}
7	−14.101 269 772 739 968 425 311 551 22	0.000
8	−14.101 269 772 739 968 425 311 551 22	0.000

[82]

输出数据表明迭代的迅速收敛性，事实上，在每一步，近似中的正确位数似乎都在成倍增长. 我们的分析将解释为什么这是正确的. ■

3.2.2 图形解释

在研究牛顿法的理论基础之前，我们给出它的一个图形解释. 从已给的描述来看，牛顿法涉及**线性化函数**. 即 f 被一个线性函数代替. 通常的做法是用 f 的泰勒级数中的前两项代替它. 因此，若

$$f(x)=f(c)+f'(c)(x-c)+\frac{1}{2!}f''(c)(x-c)^2+\cdots$$

则在 c 处的线性化产生线性函数

$$\ell(x)=f(c)+f'(c)(x-c)$$

注意 ℓ 是 f 在 c 附近的一个最佳近似，并且有 $\ell(c)=f(c)$ 以及 $\ell'(c)=f'(c)$. 因此，线性函数与 f 在点 c 处有相同的值和相同的斜率. 所以牛顿法中，我们在 r 附近的一点作 f 曲线的切线，并求出切线与 x 轴的交点(见图 3-4).

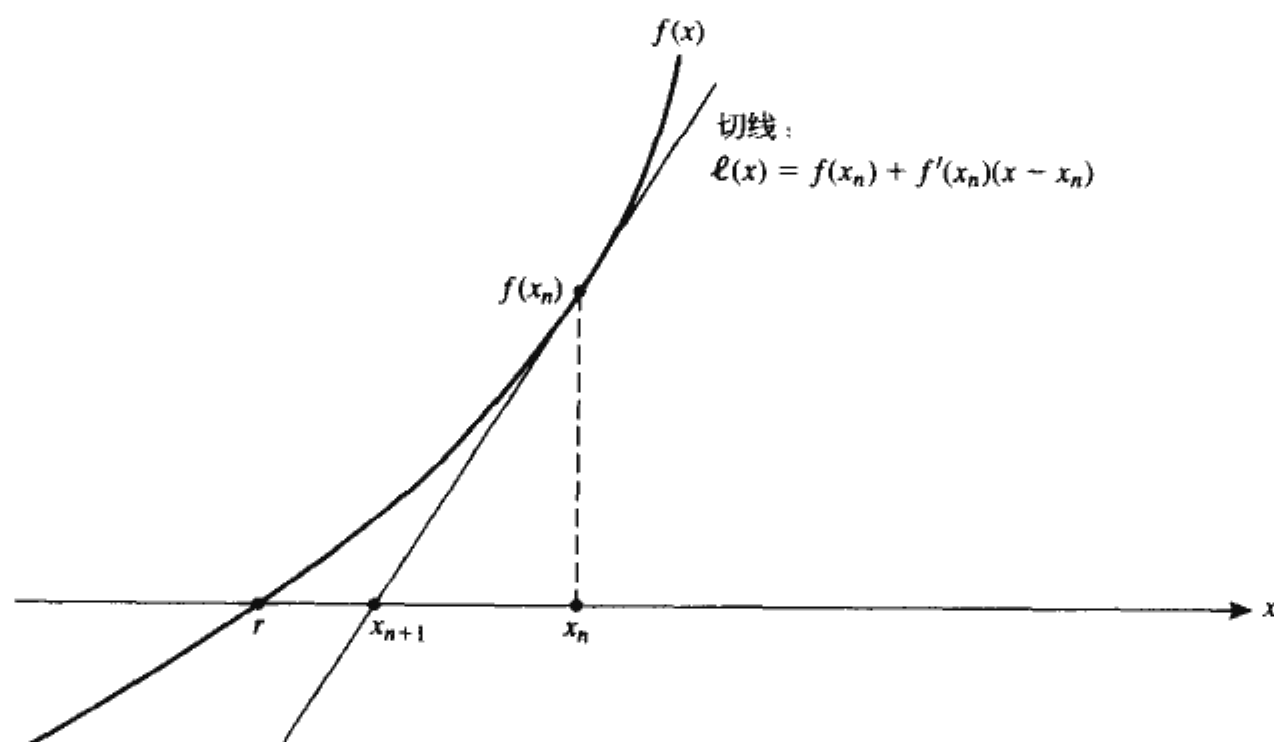


图 3-4 牛顿法的图形解释

记住图形解释，我们能容易推测出那些导致牛顿迭代失效的函数和初始点. 图 3-5 显示了这样的函数. 在此例中，曲线的形状是这样的，它对确定的初值序列[x_n]发散. 因而任何有关牛顿法的命题都必须包含一种假设，即 x_0 **充分接近于零点**或 f 的图形有指定的形状.

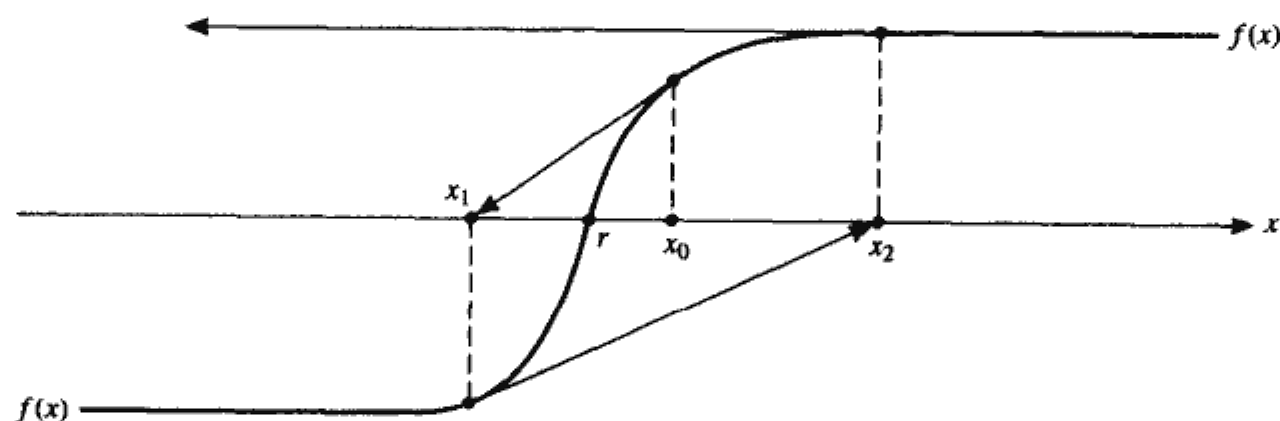


图 3-5 牛顿法不收敛的例子

3.2.3 误差分析

下面分析牛顿法中的误差. 对于**误差**，我们指的是量

$$e_n=x_n-r$$

(不考虑舍入误差.)假定 f'' 连续并且 r 是 f 的**单零点**，因此 $f(r)=0\neq f'(r)$. 从牛顿迭代的定

义，我们有

$$e_{n+1}=x_{n+1}-r=x_n-\frac{f(x_n)}{f'(x_n)}-r$$
$$=e_n-\frac{f(x_n)}{f'(x_n)}=\frac{e_nf'(x_n)-f(x_n)}{f'(x_n)} \tag{2}$$

用泰勒定理，我们有

$$0=f(r)=f(x_n-e_n)=f(x_n)-e_nf'(x_n)+\frac{1}{2}e_n^2f''(\xi_n)$$

这里ξ_n是介于x_n与r之间的一个数．重新组织此式，得

$$e_nf'(x_n)-f(x_n)=\frac{1}{2}f''(\xi_n)e_n^2$$

83~84

将其代入(2)式得

$$e_{n+1}=\frac{1}{2}\frac{f''(\xi_n)}{f'(x_n)}e_n^2\approx\frac{1}{2}\frac{f''(r)}{f'(r)}e_n^2=Ce_n^2 \tag{3}$$

假设$C\approx1$且$e_n\approx10^{-4}$．则由(3)式，我们有$e_{n+1}\approx10^{-8}$且$e_{n+2}\approx10^{-16}$．这使我们感觉只要再额外的迭代一点就能使计算精度超过机器精度！

这个等式告诉我们e_{n+1}大约为e_n^2乘以一个常数．这种令人满意的状态称为**二次收敛**．它说明在许多应用中，随着牛顿法的每次迭代，精度明显倍增．

我们还必须确定这个方法的收敛性．利用(3)式，证明的想法很简单：若e_n较小且乘数$\frac{1}{2}f''(\xi_n)/f'(x_n)$不太大，则$e_{n+1}$将比$e_n$小．定义依赖于$\delta$的量$c(\delta)$为：

$$c(\delta)=\frac{1}{2}\max_{|x-r|\leqslant\delta}|f''(x)|/\min_{|x-r|\leqslant\delta}|f'(x)| \qquad (\delta>0) \tag{4}$$

我们选择δ使其小到足以确保(4)式中的分母为正的，然后如有必要，再把δ变小使得$\delta c(\delta)<1$．这是可能的，因为当δ收敛于0时，$c(\delta)$收敛于$\frac{1}{2}|f''(r)|/|f'(r)|$，因此$\delta c(\delta)$收敛于0．固定$\delta$，令$\rho=\delta c(\delta)$．假设我们在满足$|x_0-r|\leqslant\delta$的点$x_0$处开始牛顿迭代．则$|e_0|\leqslant\delta$且$|\xi_0-r|\leqslant\delta$．所以，根据$c(\delta)$的定义，我们有

$$\frac{1}{2}|f''(\xi_0)/f'(x_0)|\leqslant c(\delta)$$

因此，(3)式产生

$$|x_1-r|=|e_1|\leqslant e_0^2c(\delta)=|e_0||e_0|c(\delta)\leqslant|e_0|\delta c(\delta)=|e_0|\rho<|e_0|\leqslant\delta$$

这说明下一个点x_1也在r的δ邻域内．因此，可以反复递推得到下列结果

$$|e_1|\leqslant\rho|e_0|$$
$$|e_2|\leqslant\rho|e_1|\leqslant\rho^2|e_0|$$
$$|e_3|\leqslant\rho|e_2|\leqslant\rho^3|e_0|$$
$$\vdots$$

就一般而言，我们有

$$|e_n|\leqslant\rho^n|e_0|$$

因为$0\leqslant\rho<1$，所以$\lim\limits_{n\to\infty}\rho^n=0$，因此$\lim\limits_{n\to\infty}e_n=0$．归纳起来，我们得到下面牛顿法的定理．

定理 1(牛顿法定理) 设 f''连续并且 r 是 f 的单零点. 则存在 r 的一个邻域和一个常数 C 使得只要牛顿法从那个邻域内的点开始，后续点就会更稳定地接近 r 并满足

$$|x_{n+1}-r|\leqslant C(x_n-r)^2 \qquad (n\geqslant 0)$$

85

在某些情况下，从任意一个初始点开始的牛顿迭代都能保证收敛. 作为例子我们给出一个这样的定理.

定理 2(凸函数的牛顿法定理) 若 $f\in C^2(\mathbb{R})$是递增的且凸的，还有零点，则这个零点是唯一的，并且从任意一个初始点开始的牛顿迭代都将收敛于此零点.

证明 回忆若对所有 x，$f''(x)>0$，则函数 f 是**凸的**. 因为 f 递增，所以在$\mathbb{R}$上，$f'>0$. 根据(3)式，$e_{n+1}>0$. 因此，对 $n\geqslant 1$，有 $x_n>r$. 又因为 f 递增，所以 $f(x_n)>f(r)=0$. 因此，根据(2)式，$e_{n+1}<e_n$. 故而，序列$[e_n]$和$[x_n]$都是递减的并都有下界(分别以 0 和 r 为下界). 因此，极限 $e^*=\lim\limits_{n\to\infty}e_n$ 和 $x^*=\lim\limits_{n\to\infty}x_n$ 存在. 从(2)式，我们有 $e^*=e^*-f(x^*)/f'(x^*)$，由此 $f(x^*)=0$ 且 $x^*=r$. ■

例 2 用牛顿法，找一个计算平方根的有效方法.

解 设 $R>0$ 且 $x=\sqrt{R}$. 则 x 是方程 $x^2-R=0$ 的根. 若我们对函数 $f(x)=x^2-R$ 使用牛顿法(1)，这个迭代公式可写成

$$x_{n+1}=\frac{1}{2}\left(x_n+\frac{R}{x_n}\right)$$

这个公式，与值域简化相结合，常常被用于计算平方根的子程序中. (这个公式是非常古老的，它归功于 Heron，他是一位生活在公元前 100 年到公元 100 年间某一时段的希腊工程师和建筑师.) 例如，当我们希望计算$\sqrt{17}$并且从 $x_0=4$ 开始. 则后续的近似值如下(以舍入形式给出，仅列出正确的数字)：

$$\begin{aligned}
x_1 &= 4.12\\
x_2 &= 4.123\,106\\
x_3 &= 4.123\,105\,625\,617\,7\\
x_4 &= 4.123\,105\,625\,617\,660\,549\,821\,409\,856
\end{aligned}$$

x_4 给出的值精确到 28 位数字，并且从这些结果中，我们观察到所期望的有效数位倍增的现象. ■

3.2.4 隐函数

牛顿法的一个有趣应用出现在求隐函数的值之中. 我们在 1.2 节中，讨论了隐函数定理，它陈述了在相当一般的条件下，方程 $G(x, y)=0$ 定义 y 作为 x 的一个函数. 若 x 给定，则可用牛顿法从方程 $G(x, y)=0$ 中解出 y. 从一个适当的初始点 y_0 开始，我们由

$$y_{k+1}=y_k-G(x,y_k)\Big/\frac{\partial G}{\partial y}(x,y_k)$$

86

定义 y_1，y_2，…. 这个方法可用来构造函数 $y(x)$的一个表. 若此表含有一个表值(x_n, y_n)，并且我们希望计算其附近表值(x_{n+1}, y_{n+1})，则由(x_{n+1}, y_n)开始进行牛顿迭代. 这一迭代结果将是使得等式 $G(x_{n+1}, y_{n+1})=0$ 成立的精确值 y_{n+1}. 因为 $G(x_n, y_n)=0$ 并且 x_{n+1}接近于

x_n，所以我们期望 $G(x_{n+1}, y_n)$ 较小并且少许几步牛顿法就能对 y_n 进行必要的校正，使等式 $G(x_{n+1}, y_{n+1})=0$ 精确成立.

例 3 建立一个 x 与 y 相对应的表，这里 y 被隐式地定义为 x 的一个函数. 利用 $G(x, y)=3x^7+2y^5-x^3+y^3-3$ 且从 $x=0$ 开始，以 0.1 的步长，依次进行到 $x=10$ 为止.

解 设 $x=0$ 且 $y=1$. 假定 4 步牛顿法将足以给出全机器精度. 在这个算法中，对 x 变量我们有 M 步，并且在每步都存在 N 次牛顿法的迭代. 根据这个算法的程序需要两个子程序或过程，其中第一个用来计算 $G(x, y)$，第二个用来计算 $\partial G/\partial y$.（在此例中，后者恰好是与 x 无关的.）这些函数是

$$G(x,y)=3x^7+2y^5-x^3+y^3-3$$
$$\frac{\partial G}{\partial y}(x,y)=10y^4+3y^2$$

下面是一个相配的算法：

```
input x←0; y←1; h←0.1; M←100; N←4
output 0, x, y, G(x, y)
for i=1 to M do
    x←x+h
    for j=1 to N do
        y←y−G(x, y)/∂G/∂y(x, y)
    end do
    output i, x, y, G(x, y)
end do
```

用上述算法计算此隐函数所得到的一些值如下：

i	x	y	$G(x, y)$
0	0.0	1.000 000	0.00
1	0.1	1.000 077	0.00
2	0.2	1.000 612	0.89×10^{-15}
⋮	⋮	⋮	⋮
20	2.0	−2.810 639	-0.82×10^{-10}
⋮	⋮	⋮	⋮
80	8.0	−19.926 35	0.56×10^{-9}
⋮	⋮	⋮	⋮
99	9.9	−26.856 18	0.12×10^{-7}
100	10.0	−27.236 85	-0.15×10^{-8}

■

87

3.2.5 非线性方程组

非线性方程组的牛顿法沿用单个方程所采用的相同策略. 因而，我们先**线性化**，然后**求解**，并根据需要，重复此步骤. 下面用一对二元方程来说明：

$$\begin{cases} f_1(x_1,x_2)=0 \\ f_2(x_1,x_2)=0 \end{cases} \tag{5}$$

假设(x_1, x_2)是(5)的近似解，我们计算校正值h_1和h_2使得(x_1+h_1, x_2+h_2)是更好的近似解. 只用二元泰勒展开式(见1.1节)中的线性项，我们有

$$\begin{cases} 0 = f_1(x_1+h_1, x_2+h_2) \approx f_1(x_1, x_2) + h_1 \dfrac{\partial f_1}{\partial x_1} + h_2 \dfrac{\partial f_1}{\partial x_2} \\ 0 = f_2(x_1+h_1, x_2+h_2) \approx f_2(x_1, x_2) + h_1 \dfrac{\partial f_2}{\partial x_1} + h_2 \dfrac{\partial f_2}{\partial x_2} \end{cases} \tag{6}$$

(6)式中的偏导在(x_1, x_2)处赋值. (6)式组成一对确定h_1和h_2的线性方程. h_1和h_2的系数矩阵是f_1和f_2的**雅可比矩阵**：

$$J = \begin{bmatrix} \partial f_1/\partial x_1 & \partial f_1/\partial x_2 \\ \partial f_2/\partial x_1 & \partial f_2/\partial x_2 \end{bmatrix}$$

为了解(6)，J必须是非奇异的. 若此条件满足，其解是

$$\begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = -J^{-1} \begin{bmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{bmatrix}$$

因此，两个二元非线性方程的牛顿法是

$$\begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \end{bmatrix} = \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \end{bmatrix} + \begin{bmatrix} h_1^{(k)} \\ h_2^{(k)} \end{bmatrix}$$

这里雅可比线性方程组

$$J \begin{bmatrix} h_1^{(k)} \\ h_2^{(k)} \end{bmatrix} = -\begin{bmatrix} f_1(x_1^{(k)}, x_2^{(k)}) \\ f_2(x_1^{(k)}, x_2^{(k)}) \end{bmatrix}$$

用高斯消元法求解——这个主题在第4章中讨论. 如果J几乎是奇异的，求解雅可比线性方程组可能就会有困难. 88

不需要新的思想来讨论包含许多变量的更大的方程组. 然而在这种情况下，矩阵向量的形式记法是非常方便的. 设$X=(x_1, x_2, \cdots, x_n)^T$，$F=(f_1, f_2, \cdots, f_n)^T$，则方程组

$$f_i(x_1, x_2, \cdots, x_n) = 0 \qquad (1 \leqslant i \leqslant n) \tag{7}$$

可简单地表成

$$F(X) = 0 \tag{8}$$

(6)式的类似表达式是

$$0 = F(X+H) \approx F(X) + F'(X)H \tag{9}$$

其中$H=(h_1, h_2, \cdots, h_n)^T$，而$F'(X)$是具有元素$\partial f_i/\partial x_j$的$n \times n$雅可比矩阵$J(X)$，即

$$F'(X) = \begin{bmatrix} \partial f_1/\partial x_1 & \partial f_1/\partial x_2 & \cdots & \partial f_1/\partial x_n \\ \partial f_2/\partial x_1 & \partial f_2/\partial x_2 & \cdots & \partial f_2/\partial x_n \\ \vdots & \vdots & \ddots & \vdots \\ \partial f_n/\partial x_1 & \partial f_n/\partial x_2 & \cdots & \partial f_n/\partial x_n \end{bmatrix}$$

校正向量H是通过解(9)中的线性方程组得到的. 理论上，这意味着

$$H = -F'(X)^{-1}F(X) \tag{10}$$

但是在实践中，H 通常是对(9)用高斯消元法确定的，从而回避了代价更高的求(10)中逆矩阵的计算．因此，对 n 个 n 元非线性方程的牛顿法是

$$X^{(k+1)} = X^{(k)} + H^{(k)} \tag{11}$$

这里雅可比方程组是

$$F'(X^{(k)})H^{(k)} = -F(X^{(k)}) \tag{12}$$

例 4 从$(1, 1, 1)^T$开始，执行6次牛顿法的迭代来求非线性方程组

$$\begin{cases} xy = z^2 + 1 \\ xyz + y^2 = x^2 + 2 \\ e^x + z = e^y + 3 \end{cases}$$

的根．

解 设

[89]

$$F(X) = \begin{bmatrix} f_1(x_1, x_2, x_3) \\ f_2(x_1, x_2, x_3) \\ f_3(x_1, x_2, x_3) \end{bmatrix} = \begin{bmatrix} x_1x_2 - x_3^2 - 1 \\ x_1x_2x_3 - x_1^2 + x_2^2 - 2 \\ e^{x_1} - e^{x_2} + x_3 - 3 \end{bmatrix}$$

求偏导，我们得到雅可比矩阵

$$F'(X) = \begin{bmatrix} x_2 & x_1 & -2x_3 \\ x_2x_3 - 2x_1 & x_1x_3 + 2x_2 & x_1x_2 \\ e^{x_1} & -e^{x_2} & 1 \end{bmatrix}$$

用初值 $X^{(0)} = (1, 1, 1)^T$ 来执行由(11)式和(12)式给出的非线性牛顿法，得到下列结果：

n	x_1	x_2	x_3
0	1.000 000 0	1.000 000 0	1.000 000 0
1	2.189 326 0	1.598 475 1	1.393 900 6
2	1.850 589 6	1.444 251 4	1.278 224 0
3	1.780 161 1	1.424 435 9	1.239 292 4
4	1.777 674 7	1.423 960 9	1.237 473 8
5	1.777 671 9	1.423 960 5	1.237 471 1
6	1.777 671 9	1.423 960 5	1.237 471 1

■

像(8)那种方程组的解常常是复杂的．关于这个主题的标准文献是 Ortega and Rheinboldt [1970]．还可以看 Rheinboldt[1974]、Ostrowski[1966]、Byrne and Hall[1973]、Schnabel, and Frank[1984]、Eaves, Gould, Peitgen and Todd[1983]以及 Allgower, Glasshoff, and Peitgen[1981]．对于较高维的牛顿法收敛性的讨论，参见 Goldstein[1966]或 Ortega and Rheinboldt[1970].

习题 3.2

1. 当牛顿法被应用到 $f(x)=\tan^{-1} x$ 时，求最小的正的初值，使得牛顿法发散．
2. 对 $f(x)=x^2-q$（其中 $q>0$）应用牛顿法．假设 $q>0.006$ 且 $k\geqslant 1$．证明若 x_n 在小数点后面有 k 个正确数字，则 x_{n+1} 在小数点后面将至少有 $2k-1$ 个正确数字．

3. 证明：设函数 f 的二阶导数 f'' 连续并且 $f(r)=0\neq f'(r)$. 若对函数 f 应用牛顿法，则 $\lim_{n\to\infty} e_{n+1}e_n^{-2}$ 存在且等于 $f''(r)/[2f'(r)]$. 另外如何把这个事实应用到程序中测试其是否二阶收敛？

4. (**Steffensen 法**)考虑迭代公式

$$x_{n+1} = x_n - f(x_n)/g(x_n)$$

其中

$$g(x) = [f(x+f(x)) - f(x)]/f(x)$$

证明在适当的假设下，这个方法是二阶收敛的.

5. 下面这个迭代公式的目的是什么？

$$x_{n+1} = 2x_n - x_n^2 y$$

确定它是某个函数的牛顿迭代.

90

6. 为了不用除法来计算倒数，我们可以通过求出函数 $f(x)=x^{-1}-R$ 零点的方法来求得 $x=1/R$. 编写一个对 f 应用牛顿法来求 $1/R$ 的短算法，在你的算法中，不能使用除法或求幂. 那么对正数 R，选择怎样的初值？

7. 定义 $x_0=0$ 和 $x_{n+1}=x_n-[(\tan x_n-1)/\sec^2 x_n]$. 在这个例子中，$\lim_{n\to\infty} x_n$ 是多少？把这个方法与牛顿法联系起来.

8. 使用计算器对多项式

$$p(x) = 4x^3 - 2x^2 + 3$$

执行 4 次牛顿法迭代，初值 $x_0=-1$.

9. 若对 $f(x)=x^3-2$ 用牛顿法，初值 $x_0=1$，试问 x_2 是多少？

10. 设计一个计算 $\sqrt[3]{R}$ 的牛顿迭代公式，这里 $R>0$. 对你的函数 $f(x)$ 做图解分析来确定使迭代收敛的初值.

11. 设计一个计算任意正实数的 5 次方根的牛顿算法.

12. 函数 $f(x)=x^2+1$ 在复平面有零点 $x=\pm i$. 对于复牛顿法，是否存在实初始点使得迭代收敛于这两个零点之一？复初始点起怎样的作用？

13. 若对 $f(x)=x^2-1$ 用牛顿法，$x_0=10^{10}$，要得到精度为 10^{-8} 的根需要迭代多少步？(用分析的方法，不要用实验方法求解.)

14. 假设 r 是函数 f 的二重根. 因此，$f(r)=f'(r)=0\neq f''(r)$. 证明若 f'' 连续，则在牛顿法中我们将有 $e_{n+1}\approx e_n/2$(线性收敛).

15. 考虑牛顿法的一种变形，在这种变形中，仅仅需要一个导数，即

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)}$$

求 C 和 s 使得

$$e_{n+1} = Ce_n^s$$

16. 证明不管选择什么(实)初始点，牛顿迭代对下列函数发散.

    a. $f(x)=x^2+1$

    b. $f(x)=7x^4+3x^2+\pi$

17. 下列序列中哪个是二次收敛的？

    a. $1/n^2$

    b. $1/2^{2^n}$

    c. $1/\sqrt{n}$

    d. $1/e^n$

    e. $1/n^n$

18. 如果初始点位于函数 f 的零点附近，为确保迭代

$$x_{n+1} = x_n - \alpha f(x_n)$$

91 线性收敛于这个零点，试求有关 α 的条件.

19. 证明：若 r 是函数 f 的 k 重零点，则通过作下列修正：

$$x_{n+1} = x_n - kf(x_n)/f'(x_n)$$

可恢复牛顿迭代的二次收敛性.

20. (续)在使用牛顿法的过程中，怎样通过分析点$(x_n, f(x_n))$的性态来发现多重零点?

21. 解方程 $f(x)=0$ 的 **Halley** 法使用了迭代公式

$$x_{n+1} = x_n - \frac{f_n f'_n}{(f'_n)^2 - (f_n f''_n)/2}$$

这里 $f_n=f(x_n)$等等. 证明当牛顿迭代应用于函数 $f/\sqrt{f'}$时，产生这个公式.

22. 从点(0, 0, 1)开始，对非线性方程组

$$\begin{cases} xy - z^2 = 1 \\ xyz - x^2 + y^2 = 2 \\ e^x - e^y + z = 3 \end{cases}$$

执行牛顿法的迭代. 请解释结果.

23. 对下列方程组执行牛顿法的两次迭代.

a. 初始点(0, 1)

$$\begin{cases} 4x_1^2 - x_2^2 = 0 \\ 4x_1x_2^2 - x_1 = 1 \end{cases}$$

b. 初始点(1, 1)

$$\begin{cases} xy^2 + x^2y + x^4 = 3 \\ x^3y^5 - 2x^5y - x^2 = -2 \end{cases}$$

计算机习题 3.2

1. 编写一个用牛顿法解方程 $x=\tan x$ 的程序. 求最接近 4.5 和 7.7 的根.
2. (续)编写且测试一个求方程 $\tan x = x$ 前10个根的程序.(这比上题要困难得多.) 注:若$\lambda_1,\lambda_2,\cdots$是这个方程的所有正根,则$\sum_{i=1}^{\infty}\lambda_i^{-2} = 1/10$.(《Amer. Math. Monthly》,10,1986,第660页.)
3. 通过使用牛顿法计算 f'的零点来求函数 $f(x)=x^{-2}\tan x$ 的最小正根.
4. 编写一个用牛顿法解方程 $x^3+3x=5x^2+7$ 的简短计算机程序. 从 $x_0=5$ 开始，计算 10 步.
5. 在 0.1 附近，方程 $2x^4+24x^3+61x^2-16x+1=0$ 有两个根. 用牛顿法求出这两个根.
6. 对本节中第一个例子，研究其根对常数项扰动的敏感性.
7. 对函数 $f(z)=z^4-1$，执行复牛顿法，初值是复平面上位于圆 $|z|<2$ 内的间距为 0.1 的网格中所有网格点. 92 把所有引起收敛于相同零点的序列的网格点涂上同一颜色. 在彩色制图计算机的终端或彩色绘图机上显示所得到的 4 色网格.
8. 编制复数运算的牛顿算法程序，并对下列函数与初始点测试此程序.

   a. $f(z)=z^2+1$, $z=3+i$

   b. $f(z)=z+\sin z-3$, $z=2-i$

   c. $f(z)=z^4+z^2+2+3i$, $z=1$
9. 编制和测试 Steffensen 法(习题 3.2.4)的程序，并用计算机习题 3.2.1～3.2.2 中的测试函数来测试这个

程序.

10. 多项式 $p(x)=x^3+94x^2-389x+294$ 有零点 1，3，-98. 因此用牛顿迭代计算两个小零点中的任一个时，点 $x_0=2$ 是一个好的初始点. 执行这个计算并且对结果作出解释.

11. 对复值函数 $f(z)=1+z^2+e^z$ 执行 5 次复牛顿法迭代，初值 $z_0=-1+4i$.

12. 求函数 $f(z)=1+z^2+e^z$ 以复数模递增来排序的前 4 个零点. 如何知道这些是前 4 个零点而你没有遗漏某个零点？

13. 对下列方程组(两个二元非线性函数)，执行 5 次牛顿法迭代

$$\begin{cases} f_1(x,y) = 1+x^2-y^2+e^x\cos y \\ f_2(x,y) = 2xy+e^x\sin y \end{cases}$$

初值为 $x_0=-1$ 和 $y_0=4$. 这个问题与上面计算机习题 11 是否有关，并且它们是否有类似的数值状态？请解释原因.

14. 用牛顿法，求下列非线性方程组的根

a. $\begin{cases} 4y^2+4y+52x=19 \\ 169x^2+3y^2+111x-10y=10 \end{cases}$

b. $\begin{cases} x+e^{-1x}+y^3=0 \\ x^2+2xy-y^2+\tan x=0 \end{cases}$

3.3 割线法

我们记得，牛顿迭代是由下式定义的：

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{1}$$

牛顿法的一个缺点是它需要求零点的函数导数. 为了克服这一缺点，人们提出了许多方法. 比如，Steffensen 迭代(习题 3.2.4).

$$x_{n+1} = x_n - \frac{[f(x_n)]^2}{f(x_n+f(x_n))-f(x_n)}$$

93

给出了一个解决这个问题的方法. 另一种方法是用**差商**

$$f'(x_n) \approx \frac{f(x_n)-f(x_{n-1})}{x_n-x_{n-1}} \tag{2}$$

代替(1)式中的 $f'(x)$. (2)式所给出的近似直接来自把 f' 作为一个极限的定义，即

$$f'(x) = \lim_{u\to x}\frac{f(x)-f(u)}{x-u}$$

如果做了这种替代，由此得到的算法称为**割线法**. 它的公式是

$$x_{n+1} = x_n - f(x_n)\left[\frac{x_n-x_{n-1}}{f(x_n)-f(x_{n-1})}\right] \qquad (n\geqslant 1) \tag{3}$$

因为 x_{n+1} 的计算需要 x_n 和 x_{n-1}. 所以开始时需要指定两个初始点. 然而，对每个新的 x_{n+1} 仅需要求一个新的 f 值. (对每个新的 x_{n+1}，Steffensen 算法则需要求两个 f 值.)

割线法的图形解释类似于牛顿法. 只是曲线的切线被割线所替代. (见图 3-6.)

例 1 用割线法求下列函数的一个零点.

$$f(x) = x^3-\sinh x+4x^2+6x+9$$

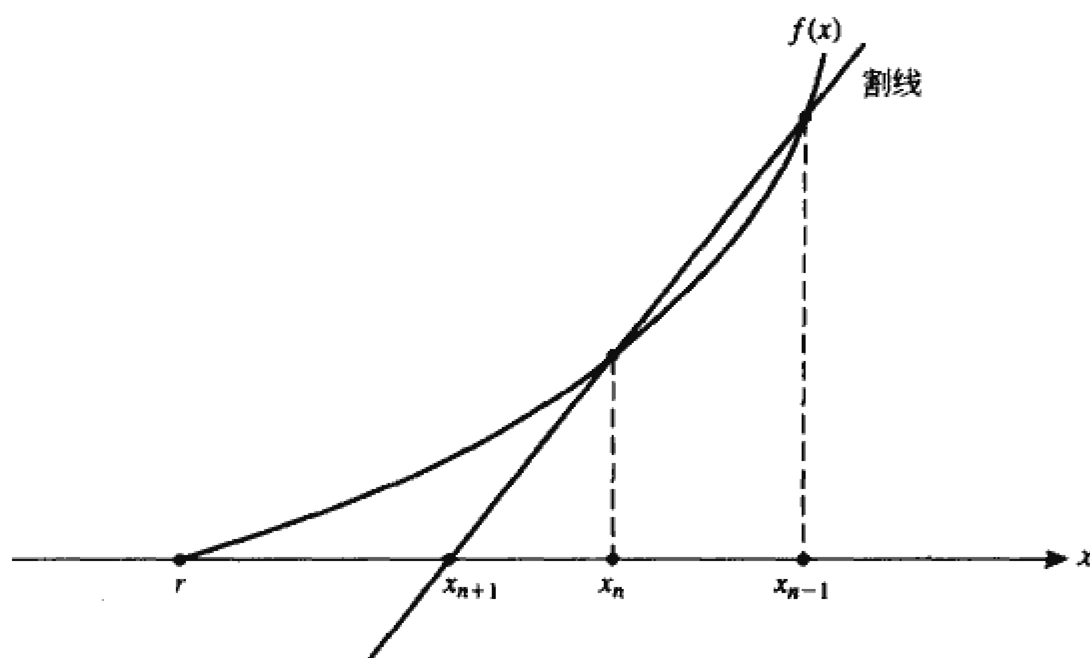


图 3-6 割线法的几何解释

解 草图提示在 7 和 8 之间存在一个零点. 取这些点作为此算法的初始点 x_0 和 x_1. 当这 [94] 个代码在类似 Marc-32 的计算机上运行时，我们就会得到如下的结果：

n	x_n	$f(x_n)$
0	8.000 00	-0.665×10^{3}
1	7.000 00	0.417×10^{2}
2	7.058 95	0.208×10^{2}
3	7.117 64	-0.183×10^{1}
4	7.112 89	0.710×10^{-1}
5	7.113 06	0.244×10^{-3}
6	7.113 06	0.191×10^{-4}

■

3.3.1 割线算法

割线法的算法可写出如下. 为了得到非递增的函数值，我们稍微对标准割线法作些修改.

```
input a, b, M, δ, ε
fa←f(a); fb←f(b)
output 0, a, fa
output 1, b, fb
for k=2 to M do
    if |fa| > |fb| then
        a↔b; fa↔fb
    end if
    s←(b−a)/(fb−fa)
    b←a
    fb←fa
    a←a−fa*s
```

```
        fa←f(a)

        output k, a, fa
        if |fa|<ε or |b−a|<δ then stop
    end do
```

注意：在伪代码中，$[a, b]$的端点可以互换以保持$|f(a)|\leqslant|f(b)|$. 因此，数对$\{x_n, x_{n-1}\}$有$|f(x_n)|\leqslant|f(x_{n-1})|$，并且下一个数对$\{x_{n+1}, x_n\}$有$|f(x_{n+1})|\leqslant|f(x_n)|$. 这确保了从第2步开始函数的绝对值就非递增.

3.3.2 误差分析

现在来分析割线法中的误差. 伪代码中包含了可能互换两个最新根的推测. 然而，在下面的分析中，我们只考虑端点不互换的简单情况.

95

记$e_n=x_n-r$，由割线法(3)的定义，我们有

$$\begin{aligned}e_{n+1} &= x_{n+1}-r=[f(x_n)x_{n-1}-f(x_{n-1})x_n]/[f(x_n)-f(x_{n-1})]-r\\ &=[f(x_n)e_{n-1}-f(x_{n-1})e_n]/[f(x_n)-f(x_{n-1})]\end{aligned}$$

提出因数$e_n e_{n-1}$并且插入$(x_n-x_{n-1})/(x_n-x_{n-1})$，我们得到

$$e_{n+1}=\left[\frac{x_n-x_{n-1}}{f(x_n)-f(x_{n-1})}\right]\left[\frac{f(x_n)/e_n-f(x_{n-1})/e_{n-1}}{x_n-x_{n-1}}\right]e_n e_{n-1} \tag{4}$$

利用泰勒定理，

$$f(x_n)=f(r+e_n)=f(r)+e_n f'(r)+\frac{1}{2}e_n^2 f''(r)+\mathcal{O}(e_n^3)$$

因为$f(r)=0$，所以由上式可得

$$f(x_n)/e_n=f'(r)+\frac{1}{2}e_n f''(r)+\mathcal{O}(e_n^2)$$

把下标改成$n-1$，得

$$f(x_{n-1})/e_{n-1}=f'(r)+\frac{1}{2}e_{n-1}f''(r)+\mathcal{O}(e_{n-1}^2)$$

将这些等式相减，得到

$$f(x_n)/e_n-f(x_{n-1})/e_{n-1}=\frac{1}{2}(e_n-e_{n-1})f''(r)+\mathcal{O}(e_{n-1}^2)$$

因为$x_n-x_{n-1}=e_n-e_{n-1}$，所以

$$\frac{f(x_n)/e_n-f(x_{n-1})/e_{n-1}}{x_n-x_{n-1}}\approx\frac{1}{2}f''(r)$$

而(4)式中第一个括号内的表达式可写成

$$\frac{x_n-x_{n-1}}{f(x_n)-f(x_{n-1})}\approx\frac{1}{f'(r)}$$

因此，我们证明了

$$e_{n+1}\approx\frac{1}{2}\frac{f''(r)}{f'(r)}e_n e_{n-1}=Ce_n e_{n-1} \tag{5}$$

此式类似于牛顿法分析中遇到的式子——3.2节中的(3)式. 为找出割线法的收敛阶，我

们要求下列渐近关系

$$|e_{n+1}| \sim A|e_n|^{\alpha} \tag{6}$$

其中 A 是正的常数．这意味着当 $n\to\infty$ 时，比率 $|e_{n+1}|/(A|e_n|^{\alpha})$ 趋向 1 并且推出 α 阶收敛．因此

96

$$|e_n| \sim A|e_{n-1}|^{\alpha} \text{ 和 } |e_{n-1}| \sim (A^{-1}|e_n|)^{1/\alpha} \tag{7}$$

在(5)式中，我们用关系式(6)和(7)中的 $|e_{n+1}|$ 和 $|e_{n-1}|$ 的渐近值代替．结果是

$$A|e_n|^{\alpha} \sim |C||e_n|A^{-1/\alpha}|e_n|^{1/\alpha}$$

此式可写成

$$A^{1+1/\alpha}|C|^{-1} \sim |e_n|^{1-\alpha+1/\alpha} \tag{8}$$

因为此关系式的左边是非零常数而 $e_n\to 0$，所以我们断定 $1-\alpha+1/\alpha=0$，或取正根，$\alpha=(1+\sqrt{5}/2)\approx 1.62$．因此，割线法的收敛率是**超线性**的(即比线性更好些)．由于关系式(8)的右边是 1，所以现在可求 A．利用等式 $1+1/\alpha=\alpha$，我们有

$$A=|C|^{1/(1+1/\alpha)}=|C|^{1/\alpha}=|C|^{\alpha-1}=|C|^{0.62}=\left|\frac{f''(r)}{2f'(r)}\right|^{0.62}$$

再用上面得到的 A，最终对割线法有

$$|e_{n+1}| \approx A|e_n|^{(1+\sqrt{5})/2}$$

因为 $(1+\sqrt{5}/2)\approx 1.62<2$，所以割线法的收敛速度不如牛顿法那样快，但是比对分法快．不过，割线法的每步只需要一次新的函数赋值，而牛顿算法中的每步却需要两次函数赋值：$f(x)$ 和 $f'(x)$．因为在这些算法中，函数赋值构成了主要的计算量，所以割线法中的两步相当于牛顿法中的一步．对于割线法中的两步，我们有

$$|e_{n+2}| \sim A|e_{n+1}|^{\alpha} \sim A^{1+\alpha}|e_n|^{\alpha^2}=A^{1+\alpha}|e_n|^{(3+\sqrt{5})/2}$$

这比牛顿法的二次收敛要好的多，因为 $(3+\sqrt{5})/2\approx 2.62$．当然，割线法中两步的工作量要多于每次迭代的工作量．

在数值分析，特别是在科学计算中，我们已讨论过的三种方法(对分法、牛顿法和割线法)说明了一种共同的现象：速度和可靠性之间的平衡．速度直接关系到计算的成本．对某些计算上比较集中的问题(比如，偏微分方程的数值解)，要求速度高于一切．而对一个被各种用户使用的通用软件而言，可靠性和稳定性是高于一切的．如果没有用户干涉，一个**稳健**的算法或子程序能够处理各种不同的数值情况．为了创作具有这两种属性的通用数学程序库，多年来人们付出了巨大的努力．其中两个杰出的范例是 IMSL[1995]和 NAG[1995]．

97

在所有科学计算任务中，除了特殊的应用外，我们不必自编程序而应该用成熟的软件包．对于求根问题，最好的软件也是相当复杂的，因为它必须结合几个过程以便既能确保整体收敛也能保证局部快速收敛．如前所述，这些目标彼此之间有些冲突．

遗憾的是，有太多的方法及其变形，不可能都在本书中描述．其中我们没有讨论的重要算法例子是由 Brent[1973]、Dekker[1969]和 Le[1985]开发的．他们把对分法和割线法的优点相结合，并且对需求根的函数除了 $f(a)f(b)\leqslant 0$ 外没有作其他任何假设．在 Brent 方法之后的思想是把对分法和割线法结合在一起并且包括一个逆二次插值法来得到一个更稳健的过程．Le

算法把对分法和使用来自目标函数值导数估计的二阶与三阶方法结合在一起. 我们鼓励感兴趣的读者去查阅上面引文以便了解这些算法的完整说明. 因为这些合成代码非常长和复杂，所有建议通过因特网来获得这些代码的软件. 我们认为只要可能，就应该使用得到确认并被检验过的软件. 通常这些精心编写和测试过的软件可免费获得. 例如，用浏览器上因特网，我们能连接 http：//gams. nist. gov，这是《Guide to Available Mathematical Software》的网址. 在那里可以找到 problem decision tree，点击进入，然后从 F. Solution of nonlinear equations 进入到 F1. Single equation 再到 F1b. Nonpolynomial. 那里你能找到 Netlib 的(求一元函数在给定域内的最小值或零点的)Brent 算法的 C 语言版本. (见附录 A.)

习题 3.3

1. 建立(4)式.
2. 在割线法中，证明若 $n \to \infty$，$x_n \to q$，并且 $f'(q) \neq 0$，则 q 是 f 的零点.
3. 用 $f(x+h)$ 和 $f(x+k)$ 的泰勒展开式导出 $f'(x)$ 的下列近似式：
$$f'(x) \approx \frac{k^2 f(x+h) - h^2 f(x+k) + (h^2 - k^2) f(x)}{(k-h)kh}$$
4. 若对函数 $f(x) = x^2 - 2$ 应用割线法，其中 $x_0 = 0$，$x_1 = 1$，试问 x_2 是多少?
5. 在割线法的一个应用中，若 $x_0 = 1$，$x_1 = 2$，$f(x_0) = 2$ 和 $f(x_1) = 5$，试问 x_2 是多少?
6. 两个序列之间的渐近相等记作 $x_n \sim y_n$ 并且表示 $\lim\limits_{n\to\infty}(x_n/y_n) = 1$. 证明若 $x_n \sim y_n$，$u_n \sim v_n$ 和 $c \neq 0$，则
   a. $cx_n \sim cy_n$
   b. $x_n^c \sim y_n^c$
   c. $u_n x_n \sim v_n y_n$
   d. 若 $y_n \sim u_n$，则 $x_n \sim v_n$
   e. $y_n \sim x_n$

98

7. 证明割线法的公式可写成下列形式：
$$x_{n+1} = \frac{f(x_n)x_{n-1} - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}$$
并解释为什么在实践中，这个公式比(3)式差.
8. (次数非常高的多项式)年金是一种货币基金，它指按固定时间间隔支付的钱(未必等值). 在年金的一种形式中，基金按每个时间间隔的固定利率 r 投资. 在每个时间间隔的末端按复利计算利息. 设支付的钱是 a_1，a_2，$\cdots$. V_i 表示年金的累加值，它在支付的钱 a_i 交纳后计算. 因而，$V_1 = a_1$ 并且
$$V_i = V_{i-1}(1+r) + a_i \qquad (i = 2,3,\cdots)$$
因子 $(1+r)$ 说明在一个时间段中 V_{i-1} 获得的利息 rV_{i-1}. 证明 $V_n = \sum\limits_{i=1}^{n} a_i x^{n-i}$，其中 $x = 1+r$. (与计算机习题 3.3.6 相关.)
9. 有两个人采取了两种不同的长达 44 年的存款方法. 第一个人前 6 年中每年存 1 000 美元，然后 38 年不提取存款以便赚得利息. 第二个人前 6 年不作任何投资，但之后每年存 1 000 美元. 然而 44 年后，两个账面的值相同. 假定这两个账户的收益是以相同的利率每年以复利计算. 试问这利率是多少? 每个账面的值又是多少?
10. 割线法伪代码中的交换是怎样影响误差分析的? 写出修正误差分析的细节.

计算机习题 3.3

1. 编写一个对函数 f 执行割线法的子程序，这里假定两个初始点是给定的．用下列函数测试此程序．
   a. $\sin(x/2)-1$
   b. $e^x-\tan x$
   c. x^3-12x^2+3x+1
2. 编程且测试割线法的一个改进，这个改进使用习题 3.3.3 中给出的 $f'(x)$ 的近似值．即，把 $f'(x)$ 的这个近似用在牛顿法的公式中．那么需要 3 个初始点．前 2 个可以是任意的．而第 3 个则用割线法得到．
3. 分别编写执行对分法、牛顿法和割线法的子程序．它们能运用于任意函数．在每种情况下，调用序列都应该包括一个用户准许的最大步数参数 M．并且用户也能够指定所需要的精度(像本节伪代码中的 ε 和 δ)．这些代码应该是单精度的．

99

   a. 用函数

$$f(x)=\tan^{-1}x-\frac{2x}{1+x^2}$$

   测试你的子程序．设法获得具有全机器精度的正零点．
   b. 用 a 中找到的零点作为对函数

$$g(x)=\tan^{-1}x$$

   应用牛顿法的初始点．
   c. 把你所编写的两个程序结合起来开发一个同时具有良好整体和局部性质的综合方法．
4. 从你计算机的共享程序库中选择一个不用导数的程序来解方程 $f(x)=0$．对下列给定区间的函数测试这个代码．
   a. $x^{20}-1$，在[0，10]上
   b. $\tan x-30x$，在[1，1.57]上
   c. $x^2-(1-x)^{10}$，在[0，1]上
   d. x^3+10^{-4}，在[−0.75，0.5]上
   e. $x^{19}+10^{-4}$，在[−0.75，0.5]上
   f. x^5，在[−1，10]上
   g. x^9，在[−1，10]上
   h. xe^{-x^2}，在[−1，4]上
   (见 Nerinckx and Haegemans[1976].)
5. 用与课本上同样的例子来编程且测试割线法．然后用 3 和 10 作为初始点重新计算．并对所发生的情况作出解释．
6. (参见习题 3.3.8)连续交纳 60 次月款到年金中．在前 5 年，支付的钱分别是 200 美元，275 美元，312 美元，380 美元和 400 美元．在交纳最后一笔支付款后，年金的累积值是 24 738 美元．试问月利率是多少？用割线法求所出现的多项式的零点．

3.4 不动点和函数迭代

牛顿法和 Steffensen 法是通过形如

$$x_{n+1}=F(x_n)\qquad(n\geqslant 0)\tag{1}$$

的公式来计算点序列过程的一些实例．由这样的等式来定义的算法称为**函数迭代**．在牛顿法

中，函数 F 为

$$F(x) = x - \frac{f(x)}{f'(x)}$$

100

而在 Steffensen 法中，有

$$F(x) = x - \frac{[f(x)]^2}{f(x+f(x)) - f(x)}$$

我们还可以列举出许多其他函数迭代通用过程的例子，并将概要地研究这个一般的理论.

公式(1)可用于生成不收敛的序列——例如，序列 1，3，9，27，…，这个序列是在 $x_0=1$ 和 $F(x)=3x$ 的情况下产生的. 然而，我们的兴趣主要是在那些$\lim\limits_{n\to\infty} x_n$ 存在的情况. 因而假设

$$\lim_{n\to\infty} x_n = s$$

那么 s 和 F 之间存在什么关系呢？如果 F 连续，那么

$$F(s) = F(\lim_{n\to\infty} x_n) = \lim_{n\to\infty} F(x_n) = \lim_{n\to\infty} x_{n+1} = s$$

因此，$F(s)=s$，并且我们称 s 为函数 F 的**不动点**. 我们可以认为不动点是函数在迭代过程中"锁定"的值.

一个数学问题常常能归结成为一个求函数不动点的问题. 一些非常有意思的应用出现在微分方程、最优化理论和其他领域之中. 通常需寻找不动点的函数 F 是一个从一个向量空间到另一个向量空间内的映射. 我们打算分析 F 把某个闭集 $C\subseteq\mathbb{R}$ 映射到其自身内这种最简单的情况，而其所证明的定理与**压缩映射**有关. 如果存在一个小于 1 的数 λ，使得对 F 定义域中所有的点 x 和 y，有

$$|F(x) - F(y)| \leqslant \lambda |x - y| \tag{2}$$

这个映射(或函数)F 被称为是**压缩的**. 如图 3-7 所示，压缩函数 F 把 x 和 y 间的距离映射成 $F(x)$和 $F(y)$间的较短距离.

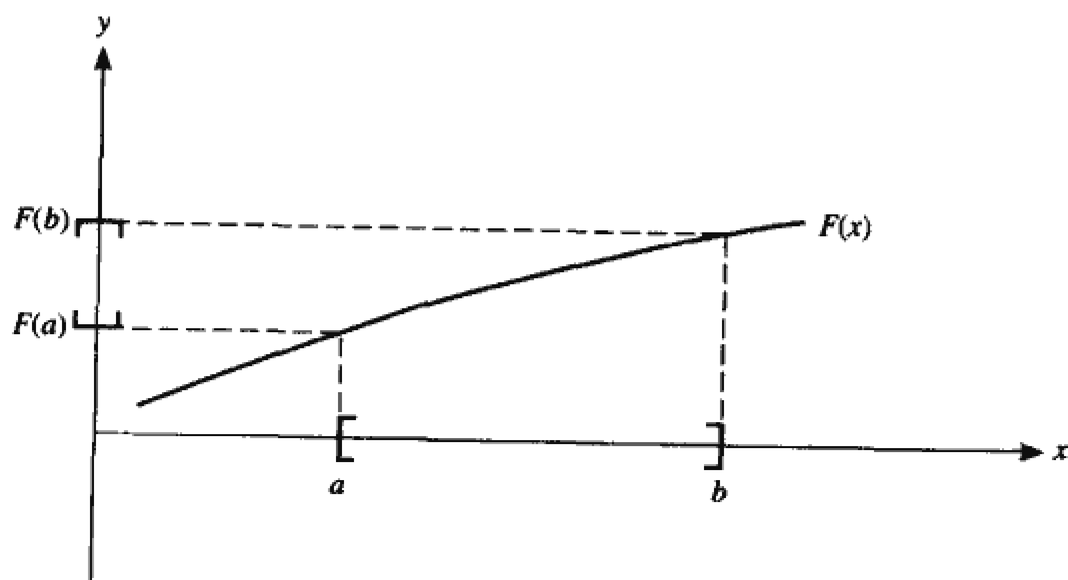


图 3-7 压缩映射图解的例子

101

定理 1(压缩映射定理) *设 C 是实轴的一个闭子集. 若 F 是一个 C 到 C 内的压缩映射，则 F 有唯一的不动点. 而这个不动点是初始点为 $x_0\in C$ 的由(1)式所产生的每个序列的极限.*

证明 利用压缩性质(2)和(1)式得

$$|x_n - x_{n-1}| = |F(x_{n-1}) - F(x_{n-2})| \leqslant \lambda |x_{n-1} - x_{n-2}| \tag{3}$$

重复这个论证，得到

$$|x_n - x_{n-1}| \leqslant \lambda |x_{n-1} - x_{n-2}| \leqslant \lambda^2 |x_{n-2} - x_{n-3}| \leqslant \cdots \leqslant \lambda^{n-1} |x_1 - x_0|$$

因为 x_n 可写成如下形式

$$x_n = x_0 + (x_1 - x_0) + (x_2 - x_1) + \cdots + (x_n - x_{n-1})$$

所以我们看到序列[x_n]收敛当且仅当级数

$$\sum_{n=1}^{\infty} (x_n - x_{n-1})$$

收敛. 要证明这个级数收敛，只需证明级数

$$\sum_{n=1}^{\infty} |x_n - x_{n-1}|$$

收敛即可. 这很容易证明，因为我们可用比较法和前面的工作：

$$\sum_{n=1}^{\infty} |x_n - x_{n-1}| \leqslant \sum_{n=1}^{\infty} \lambda^{n-1} |x_1 - x_0| = \frac{1}{1-\lambda} |x_1 - x_0|$$

由于此序列收敛，所以设 $s = \lim_{n\to\infty} x_n$. 因而，用前面的记号，有 $F(s) = s$.（显然压缩性蕴涵 F 的连续性.）下面证明不动点的唯一性. 如果 x 和 y 都是不动点，则

$$|x - y| = |F(x) - F(y)| \leqslant \lambda |x - y|$$

因为 $\lambda < 1$，所以 $|x - y| = 0$. 最后，因为 s 是 C 中序列的极限，则所得到的点 s 属于 C. ■

对所有完备度量空间到其自身内的压缩映射，定理 1 都成立.

[102] **例 1** 证明压缩映射定理中序列[x_n]满足**柯西收敛准则**.

解 回忆关于序列[x_n]的柯西准则：给定任意的 ε，存在一个整数 N 使得当 n，$m \geqslant N$ 时，有 $|x_n - x_m| < \varepsilon$. 若 $n \geqslant m \geqslant N$，则根据三角不等式和(3)后面的式子，有

$$\begin{aligned}
|x_n - x_m| &\leqslant |x_n - x_{n-1}| + |x_{n-1} - x_{n-2}| + \cdots + |x_{m+1} - x_m| \\
&\leqslant \lambda^{n-1} |x_1 - x_0| + \lambda^{n-2} |x_1 - x_0| + \cdots + \lambda^m |x_1 - x_0| \\
&= \lambda^m |x_1 - x_0| (1 + \lambda + \lambda^2 + \cdots + \lambda^{n-1-m}) \\
&\leqslant \lambda^N |x_1 - x_0| (1 + \lambda + \lambda^2 + \cdots) \\
&= \lambda^N |x_1 - x_0| (1 - \lambda)^{-1}
\end{aligned}$$

为了对任意 $\varepsilon > 0$，存在一个 N 使得当 n，$m \geqslant N$ 时，有 $|x_n - x_m| < \varepsilon$. 我们只要增加 N 直到 $\lambda^N |x_1 - x_0| (1-\lambda)^{-1} < \varepsilon$ 即可. ■

例 2 证明如下递归定义的序列[x_n]收敛.

$$\begin{cases} x_0 = -15 \\ x_{n+1} = 3 - \dfrac{1}{2} |x_n| \qquad (n \geqslant 0) \end{cases}$$

解 函数 $F(x) = 3 - |x|/2$ 是压缩的，因为根据三角不等式，有

$$|F(x) - F(y)| = \left| 3 - \frac{1}{2}|x| - 3 + \frac{1}{2}\Big| |y| - |x| \Big| \right| = \frac{1}{2}\Big| |y| - |x| \Big| \leqslant \frac{1}{2} |y - x|$$

根据定理 1，上述序列必收敛于 F 的唯一不动点，容易看出它是 2. 在此定理中，我们可取 C 为$\mathbb{R}$. ■

例 3 用压缩映射定理来计算下列函数的不动点

$$F(x) = 4 + \frac{1}{3}\sin 2x$$

解 根据中值定理，我们有

$$|F(x) - F(y)| = \frac{1}{3}|\sin 2x - \sin 2y| = \frac{2}{3}|\cos 2\zeta||x - y| \leqslant \frac{2}{3}|x - y|$$

其中 ζ 介于 x 与 y 之间. 这说明 F 是压缩的，并且 $\lambda=2/3$. 根据定理 1，F 有不动点. 求此不动点的计算机程序是基于下面这个算法的，它从初值 4 开始，执行 20 次迭代：

```
input x ← 4;  M ← 20
for k = 1 to M do
    x ← 4 + (1/3) sin2x
    output k, x
end do
```

此程序产生 20 行输出数据，下面显示其中的若干行：

[103]

k	x
1	4.329 786 1
2	4.230 895 1
3	4.273 633 8
⋮	⋮
14	4.261 483 0
15	4.261 484 0
16	4.261 483 6
⋮	⋮
20	4.261 483 7

在最后一行中，给出了 7 位精确小数的不动点. ■

误差分析

现在让我们来分析函数迭代过程中的误差. 假设 F 有不动点 s，而序列$[x_n]$由公式 $x_{n+1} = F(x_n)$定义. 设

$$e_n = x_n - s$$

若 F'存在且连续，则根据中值定理，

$$x_{n+1} - s = F(x_n) - F(s) = F'(\zeta_n)(x_n - s)$$

或

$$e_{n+1} = F'(\zeta_n)\, e_n$$

这里点 ζ_n 介于 x_n 与 s 之间. 对所有的 x，条件 $|F'(x)|<1$ 保证误差数量递减. 若 e_n 较小，则 ζ_n 就接近 s，且 $F'(\zeta_n)\approx F'(s)$. 当 $F'(s)$较小时，人们预料会迅速收敛. 一个理想的情况是 $F'(s)=0$. 此时，将额外使用泰勒级数中的一项. 为了同时处理几种情况，假设 q 是一个整数使得

$$F^{(k)}(s)=0,1\leqslant k<q\text{ ,但 }F^{(q)}(s)\neq 0$$

根据 $F(x_n)$在 s 附近展开的泰勒级数，我们有

$$\begin{aligned}e_{n+1}&=x_{n+1}-s=F(x_n)-F(s)\\&=F(s+e_n)-F(s)\\&=[F(s)+e_nF'(s)+\frac{1}{2}e_n^2F''(s)+\cdots]-F(s)\\&=e_nF'(s)+\frac{1}{2}e_n^2F''(s)+\cdots+\frac{1}{(q-1)!}e_n^{q-1}F^{(q-1)}(s)+\frac{1}{q!}e_n^qF^{(q)}(\zeta_n)\end{aligned}$$

[104]

因此

$$e_{n+1}=\frac{1}{q!}e_n^qF^{(q)}(\zeta_n) \tag{4}$$

如果我们知道$\lim\limits_{n\to\infty}x_n=s$，那么(4)式蕴涵

$$\lim_{n\to\infty}\frac{|e_{n+1}|}{|e_n|^q}=\frac{1}{q!}F^{(q)}(s) \tag{5}$$

回忆在 1.2 节中，对任何收敛于点 s 的序列$[x_n]$(不管它是否由函数迭代引起)，我们定义**收敛阶**是使得极限

$$\lim_{n\to\infty}\frac{|x_{n+1}-s|}{|x_n-s|^q}$$

存在且非零的最大实数 q. 这样的 q 不一定存在，而当它存在时，也不一定是一个整数.

例如，若 $F'(s)=0$ 且 $F''(s)\neq 0$，则 $q=2$ 并且有

$$e_{n+1}=\frac{1}{2}e_n^2F''(\zeta_n)$$

这使人联想到牛顿法在 3.2 节中的(3)式. 事实上，牛顿法使用

$$F(x)=x-\frac{f(x)}{f'(x)}$$

并且对它，我们有

$$F'(x)=1-\frac{f'(x)f'(x)-f(x)f''(x)}{[f'(x)]^2}=\frac{f(x)f''(x)}{[f'(x)]^2}$$

因为 F 的不动点是 f 的零点，所以 $F'(s)=0$. 此外，

$$F''(x)=\frac{[f'(x)]^2[f(x)f'''(x)+f''(x)f'(x)]-[f(x)f''(x)][2f'(x)f''(x)]}{[f'(x)]^4}$$

因此，通常我们有

$$F''(s)=\frac{f''(s)}{f'(s)}\neq 0$$

若 F 用在不动点 s 附近的泰勒级数来表示

$$\sum_{k=0}^{\infty} \frac{1}{k!} F^{(k)}(s)(x-s)^k$$

并且若 $x_{n+1}=F(x_n)$，则收敛阶是使得 $F^{(q)}(s)\neq 0$ 的第一个整数 q. 习题 3.4.16 要求给出这个命题的证明.

105

习题 3.4

1. 若 F 是从$[a, b]$到$[a, b]$的压缩并且 $x_{n+1}=F(x_n)$，$x_0\in[a, b]$，则对适当的 C，有 $|x_n-s|\leqslant C\lambda^n$. 证明这个结论并且给出 C 的一个上界. 这里 s 是 F 的不动点.
2. 证明：若 F：$[a, b]\to\mathbb{R}$，F'连续，且在$[a, b]$上 $|F'(x)|<1$，则 F 是压缩的. 试问 F 是否一定有不动点?
3. 证明：若 F 是$[a, b]$到$[a, b]$内的连续映射，则 F 必有一个不动点. 然后确定这个命题对从$\mathbb{R}$到$\mathbb{R}$的函数是否成立?
4. 证明下列函数在指定的区间上是压缩的. 并求(2)式中最佳的 λ 值.
   a. $(1+x^2)^{-1}$ 在任何区间上
   b. $x/2$ 在 $1\leqslant x\leqslant 5$ 上
   c. $\tan^{-1} x$ 在任何不含 0 的闭区间上
   d. $|x|^{\frac{3}{2}}$ 在 $|x|\leqslant 1/3$ 上
5. 天文学中的开普勒方程是 $x=y-\varepsilon\sin y$，$0<\varepsilon<1$. 证明对每个 $x\in[0, \pi]$，存在一个 y 满足这个方程. 并把这解释为一个不动点的问题.
6. 考虑形如 $F(x)=x+f(x)g(x)$的迭代函数，这里 $f(r)=0$ 且 $f'(r)\neq 0$. 试求关于函数 g 的精确条件使得当在 r 附近开始时，这个函数迭代法会**三次收敛**于 r.
7. 若你把一个数输入袖珍计算器中，然后反复按 cosine 键，最终会出现怎样的数? 请提供一个证明.
8. 证明：若在区间$[x_0-\rho, x_0+\rho]$上，$|F'(x)|\leqslant\lambda<1$，其中 $\rho=|F(x_0)-x_0|/(1-\lambda)$，则由迭代 $x_{n+1}=F(x_n)$生成的序列收敛.
9. 如果应用于函数 f 的牛顿法三次收敛于 f 的零点，那么 f 必须具有怎样的特殊性质?
10. 如果我们试图通过把牛顿法应用于方程 $F(x)-x=0$ 来求 F 的不动点，试问会产生怎样的迭代公式?
11. 证明：若 f'在$[a, b]$上是连续且正的，并且 $f(a)f(b)<0$，则 f 在(a, b)内只有一个零点. 同时证明用适当的参数 λ 对 $F(x)=x+\lambda f(x)$，应用函数迭代法就可得到这个零点.
12. 设 p 是一个正数. 下列表达式的值是多少?

$$x=\sqrt{p+\sqrt{p+\sqrt{p+\cdots}}}$$

   注意：这个表达式可解释为 $x=\lim\limits_{n\to\infty}x_n$，$x_1=\sqrt{p}$，$x_2=\sqrt{p+\sqrt{p}}$，等等. 提示：观察 $x_{n+1}=\sqrt{p+x_n}$.
13. (续)设 $p>1$. 下列连分式的值是多少?

$$x=\cfrac{1}{p+\cfrac{1}{p+\cfrac{1}{p+\cdots}}}$$

   用上题的思想来解这个问题. 并用压缩映射定理来证明数值序列收敛.

106

14. (续)提出一个求二次方程 $x^2+px+q=0$ 根的迭代法.
15. 设 F 是区间$[a, b]$到其自身内的一个压缩映射，而 s 是 F 的不动点. 如果 $a\leqslant x\leqslant b$ 并且 $|F(x)-x|<\varepsilon$，那么是否可推出有 $|x-s|<\varepsilon$? 证明 $|x-s|<\varepsilon(1-\lambda)^{-1}$，这里 λ 是(2)式中的常数.

16. 证明课本中本节结尾处有关函数迭代法收敛阶的命题.

17. 大多数迭代过程不像由 $x_{n+1}=F(x_n)$，F：$\mathbb{R}\to\mathbb{R}$ 所表达的那样简单. 例如，可能有函数 F：$\mathbb{R}^2\to\mathbb{R}^2$. 证明对分法和割线法就是这种类型. 并明确地定义每种情况中的 F.

18. 证明：若 F' 连续并且在区间 $[a, b]$ 上 $|F'(x)|<1$，则 F 在 $[a, b]$ 上是压缩的. 请说明对于开区间，这个结论未必成立.

19. 如果函数迭代法应用于函数 $F(x)=2+(x-2)^4$ 上，并从 $x=2.5$ 开始，试问其收敛阶是多少？求出使这个函数迭代收敛的初值范围. 注意：2 是不动点.

20. 证明下列函数在所给的定义域上是压缩的，但是它们在这些定义域上却没有不动点. 为什么这与压缩映射定理不矛盾？

    a. $F(x)=3-x^2$ 在 $[-1/4, 1/4]$ 上

    b. $F(x)=-x/2$ 在 $[-2, -1]\cup[1, 2]$ 上

21. 证明：若 f 在 $[a, b]$ 上连续，且满足 $a\leqslant f(a)$ 且 $f(b)\leqslant b$，则 f 在区间 $[a, b]$ 内有不动点. 注意：我们没有假设 $a\leqslant f(x)\leqslant b$，$x\in[a, b]$.

22. 在区间 $[c, d]$ 上增加什么样最低条件才能使得 $[a, b]$ 到 $[c, d]$ 内的每个连续映射都有不动点？

23. 求下列序列的收敛阶.

    a. $x_n=(1/n)^{\frac{1}{2}}$

    b. $x_n=\sqrt[n]{n}$

    c. $x_n=(1+1/n)^{\frac{1}{2}}$

    d. $x_{n+1}=\tan^{-1}x_n$

24. 为了求函数 f 的零点，我们可求函数 $F(x)=x-f(x)/f'(x)$ 的不动点. 为了求 F 的不动点，我们可用牛顿法解 $F(x)-x=0$. 当如此这般操作时，生成序列 x_n 的公式是什么？

25. 证明由 $F(x)=4x(1-x)$ 定义的函数 F 把区间 $[0, 1]$ 映射到其自身内，并且 F 不是压缩的. 再证明它有不动点. 而且为什么这与压缩映射定理不矛盾？

26. 若对 $f(x)=(1+x^2)^{-1}/2$ 用函数迭代法，并且从 $x_0=7$ 开始. 试问由此产生的序列是否收敛？如果收敛，那么极限是什么？细致地给出你的回答.

27. 证明或否定：若 F：$\mathbb{R}\to[a, b]$，且 F 在 $[a, b]$ 上是压缩的，则 F 有唯一的不动点，它可以通过函数迭代法来获得，迭代初值可以是任何实值.

28. 举出没有不动点但有下列特性的函数实例：

    [107]

    a. f：$[0, 1]\to[0, 1]$

    b. f：$(0, 1)\to(0, 1)$ 且连续

    c. f：$A\to A$ 且连续，其中 $A=[0, 1]\cup[2, 3]$

    d. f：$\mathbb{R}\to\mathbb{R}$ 且连续

29. 证明函数 $f(x)=2+x-\tan^{-1}x$ 有性质 $|f'(x)|<1$. 再证明 f 没有不动点. 并解释为什么这与压缩映射定理不矛盾.

30. 这个问题涉及函数 $F(x)=10-2x$. 证明 F 有不动点. 设 x_0 是任意的，且定义 $x_{n+1}=F(x_n)$，$n\geqslant 0$. 求 x_n 的非递归公式. 证明除非 x_0 是一个给定的特殊值，否则这个函数迭代法不会产生收敛序列. x_0 的这个特殊值是多少？为什么这与压缩映射定理不矛盾？

31. 设 F 在一个开区间内是连续可微的，并且假设 F 在这个开区间内有不动点 s. 证明：若 $|F'(s)|<1$，则只要初始点充分接近 s，由函数迭代法定义的序列就会收敛于 s. 提示：选择 λ 使得 $|F'(r)|<\lambda<1$，然后，考虑中点在 r 的区间，并且在此区间内有 $|F'(x)|<1$.

32. 设 $1/2\leqslant q\leqslant 1$，并且定义 $F(x)=2x-qx^2$. 试问在怎样的区间上能保证用 F 的迭代法收敛于不动点？(这个问题与习题 3.2.5 相关.)
33. 写出两种求函数 $f(x)=2x^2+6e^{-x}-4$ 零点的不动点的过程.
34. 在$[1/2, \infty]$, $[1/8, 1]$, $[1/4, 2]$, $[0, 1]$, $[1/5, 3/2]$中的哪个区间上，函数 $f(x)=\sqrt{x}$是压缩的？
35. 一个函数被称为**多重收缩的**，如果

$$|F(F(x))-F(x)|\leqslant \lambda |F(x)-x| \qquad (\lambda<1)$$

证明每个收缩都是多重收缩的，但是多重收缩未必是收缩的，也不一定是连续的.
36. 如果函数迭代法用于 $F(x)=x^2+x-2$，并且产生一个收敛的正数序列，那么这个序列的极限是什么？初始点是什么？
37. 考虑形如 $F(x)=x-f(x)/f'(x)$的函数，其中 $f(r)=0$ 且 $f'(r)\neq 0$. 求关于函数 f 的精确条件使得在 r 附近开始迭代时，这个函数迭代法至少**三次**收敛于 r.
38. 分析作为函数迭代实例的 Steffensen 法(习题 3.2.4). 确定其收敛阶.
39. 一名学生记错了牛顿法，并写成 $x_{n+1}=f(x_n)/f'(x_n)$. 问这个方法是否能求出 f 的零点？收敛阶又是多少？
40. 证明下列计算$\sqrt{R}$的方法有三次收敛性：

$$x_{n+1}=\frac{x_n(x_n^2+3R)}{3x_n^2+R}$$

41. 考虑形如 $x_{n+1}=x_n-f(x_n)/g(x_n)$的迭代方法. 假设它收敛于点 ξ，这个点是函数 f 的单零点但不是函数 g 的零点. 请建立 f 和 g 之间的关系使得这个方法的收敛阶是 3 或更高.

108

3.5 求多项式的根

在前面各节中所讨论的方法(特别是牛顿法)当然能用于多项式. 但是在求多项式的根时，如果可能，我们就应该充分利用这种函数的特殊结构. 此外，多项式问题时常因为我们希望计算已知多项式的复根(即使多项式有实系数)或所有根而变得复杂起来. 因而，求多项式根的课题需要特殊处理，的确，它也受到这种待遇已有差不多 400 年了.

我们从某些重要的理论结论开始，读者可能已熟悉这些结论中的大多数. 我们把多项式写成如下形式

$$p(z)=a_nz^n+a_{n-1}z^{n-1}+\cdots+a_2z^2+a_1z+a_0 \tag{1}$$

其中，系数 a_k 和变量 z 可能是复数. 若 $a_n\neq 0$，则 p 为 ***n* 次**多项式. 我们对求 p 的根感兴趣. 那么是否存在根？下面的定理(首先由高斯于 1799 年证明)回答了这个问题. 这里给出的来自 1938 年的证明是 G. H. Hardy[1960，附录一]和 Fefferman[1967]所写的.

定理 1(代数基本定理) *每个非常数多项式在复数域内至少有一个根.*

证明 设 p 是非常数多项式. 我们希望证明对某个 $z_0\in\mathbb{C}$，$p(z_0)=0$. 因为 p 不是常数，所以当 $|z|\to\infty$时，$|p(z)|\to\infty$. 设 D 是一个中心在 0 的圆盘且在圆盘外面 $|p(z)|\geqslant|p(0)|$. 设 z_0 是一个达到$\inf\limits_{z\in D}|p(z)|$的点. 因为$0\in D$，所以 $|p(z_0)|\leqslant|p(0)|$. 因此对所有 $z\in\mathbb{C}$，$|p(z_0)|\leqslant|p(z)|$. 令 $q(z)=p(z+z_0)$. 我们希望证明 $q(0)=0$，从而 $p(z_0)=0$. 记 $q(z)=c_0+c_jz^j+\cdots+c_nz^n=c_0+c_jz^j+z^{j+1}r(z)$，其中 $c_j\neq 0$ 且 r 是一个多项式

(可能是 0). 现在我们要证明 $c_0=0$. 假如 $c_0\neq 0$. 设 w 是任意一个满足 $c_jw^j=-c_0$ 的复数. 定义 $N=\sup_{0<\varepsilon<1}|r(\varepsilon w)|$. 选择足够小的 $\varepsilon\in(0,1)$使得 $\varepsilon|w|^{j+1}N<|c_0|$. 因此，我们得到如下的矛盾：

$$\begin{aligned}|q(\varepsilon w)| &\leqslant |c_0+c_j\varepsilon^j w^j|+\varepsilon^{j+1}|w^{j+1}||r(\varepsilon w)|\\ &=|c_0-c_0\varepsilon^j|+\varepsilon^j\varepsilon|w|^{j+1}N\\ &<|c_0|(1-\varepsilon^j)+\varepsilon^j|c_0|=|c_0|=|q(0)|\\ &=|p(z_0)|\leqslant|p(z_0+\varepsilon w)|=|q(\varepsilon w)|\end{aligned}$$ ■

这个定理及其证明的历史可阅读 Kline[1972，第 597～606 页]. 由刘维尔定理的推论可得到这个现代的证法. 例如，见 Bak and Newman[1982]，Henrici[1974]，Ahlfors[1966]，或其他有关复分析的教科书.

代数基本定理没有断言实根的存在性，而且像 x^2+1 这样最简单的例子，也表明即使多项式的系数都是实数，它也未必有实根.

109

若次数 n 至少为 1 的多项式 p 被线性因式 $z-c$ 除，则结果产生商 q 和余项 r. 后者是一个复数，而前者是一个 $n-1$ 次多项式. 我们可用等式

$$p(z)=(z-c)q(z)+r$$

表示这个过程. 由此我们看到(通过令 $z=c$) $p(c)=r$. 这个事实称为**剩余定理**. 若 c 是 p 的根，则 $r=0$，并且有

$$p(z)=(z-c)q(z)$$

因此，$z-c$ 是 $p(z)$的一个因式. 这个结论称为**因子定理**.

记 $p(z)=(z-r_1)q_1(z)$，其中 r_1 是 p 的任何一个根. 根据代数基本定理，q_1 有一个根，比如说是 r_2(除非 q_1 是 0 次). 因此，$q_1(z)$有因式 $z-r_2$，并且我们可记 $p(z)=(z-r_1)(z-r_2)q_2(z)$. 依此进行下去，最后到达终止点，由于在每一步，相继的 $q_k(z)$的次数减 1. 因此，q_n 将是一个常数，而最后的等式是

$$p(z)=(z-r_1)(z-r_2)\cdots(z-r_n)q_n$$

这证明了 n 次多项式 p 有一个 n 个线性因式乘积的分解式，并且每个因式对应于 p 的一个根. 显然 p 不可能有其他的根，因为如果 z 是任何不同于 $r_1,r_2,\cdots,r_n$ 的复数，那么积 $\prod_{k=1}^{n}(z-r_k)$ 就不可能是 0. 因为有些根 r_k 可能彼此相等，所以 n 次多项式至多有 n 个根. 根的**重数**是指对应于此根的因式在 p 的因式分解中出现的次数. 把前面的注释与代数基本定理相结合，我们就获得了下列定理：

定理 2(多项式复根定理) 一个 n 次多项式在复平面内恰好有 n 个根，每个根的重复次数等于它的重数.

通常我们想要知道多项式的根在复平面内所处的大概位置. 文献中，有许多诸如笛卡儿正负号法则这样的有关**根定位**问题的结论. 例如，见 Young and Gregory[1972]的 5.5 节，Stoer and Bulirsch[1980]的 5.5 节，Marden[1966]的论文丛集，或 Henrici[1974]的 3 卷论文集中的

第 1 卷. 下列定理取自 Conte and de Boor[1980]教科书中，它提供了关于根的模的一个容易计算的上界.

定理 3(定位定理) 等式(1)中多项式的所有根位于中心是复平面原点的开圆盘内，并且它的半径是

$$\rho = 1 + |a_n|^{-1} \max_{0 \leqslant k < n} |a_k|$$

110

证明 令 $c = \max_{0 \leqslant k \leqslant n} |a_k|$，因此 $c|a_n|^{-1} = \rho - 1$. 如果 $c=0$，则结论自然成立. 因此假定 $c>0$. 于是 $\rho>1$. 若 $|z| \geqslant \rho$，则(因为 $\rho>1$)

$$\begin{aligned}
|p(z)| &\geqslant |a_n z^n| - |a_{n-1}z^{n-1} + \cdots + a_0| \\
&\geqslant |a_n z^n| - c\sum_{k=0}^{n-1} |z|^k \\
&> |a_n z^n| - c|z|^n(|z|-1)^{-1} \\
&= |a_n z^n| \{1 - c|a_n|^{-1}(|z|-1)^{-1}\} \\
&\geqslant |a_n z^n| \{1 - c|a_n|^{-1}(\rho-1)^{-1}\} = 0
\end{aligned}$$

■

例 1 求中心在原点的圆盘，使得它包含多项式

$$p(z) = z^4 - 4z^3 + 7z^2 - 5z - 2$$

的所有根.

解 根据定理 3，这样一个圆盘的半径是

$$\rho = 1 + |a_4|^{-1} \max_{0 \leqslant k < 4} |a_k| = 8$$

■

另一个分析多项式的有用思想是：取(1)式的多项式 p，并考虑函数 $s(z) = z^n p(1/z)$. 因此，

$$\begin{aligned}
s(z) &= z^n\left[a_n\left(\frac{1}{z}\right)^n + a_{n-1}\left(\frac{1}{z}\right)^{n-1} + \cdots + a_0\right] \\
&= a_n + a_{n-1}z + a_{n-2}z^2 + \cdots + a_0 z^n
\end{aligned}$$

这说明 s 是一个次数至多为 n 的多项式. 它的系数可以立刻写出，因为除了次序颠倒外，它们与 p 的系数相同. 现在，显然对非零复数 z_0，条件 $p(z_0)=0$ 等价于条件 $s(1/z_0)=0$. 因此，我们有下列结论.

定理 4(根定位定理) 若 s 的所有根都在圆盘 $\{z : |z| \leqslant \rho\}$ 内，则 p 的所有非零根都在圆盘 $\{z : |z| < \rho^{-1}\}$ 外.

例 2 求一个中心在原点的圆盘，使得它不含 p 的根，这里 p 是例 1 中的多项式.

解 定理 4 中提到的多项式 s 是

$$s(z) = -2z^4 - 5z^3 + 7z^2 - 4z + 1$$

因而，根据定理 3，s 的所有根都在中心在原点，半径为 $\rho = 1 + |a_4|^{-1}\max_{0 \leqslant k < 4} |a_k| = 9/2$ 的圆盘内. 根据定理 4，p 的所有根位于半径是 2/9 的圆盘外. 因此，p 的所有根就在复平面上的环 $2/9 < |z| < 8$ 内.

■ 111

3.5.1 霍纳算法

为了高效率地计算多项式的值，我们需要**霍纳算法**．这个算法也称为**嵌套乘法**和**综合除法**．它对其他课题也是很有用的．如果给定多项式 p 和复数 z_0，那么霍纳算法将产生数 $p(z_0)$ 和多项式

$$q(z)=\frac{p(z)-p(z_0)}{z-z_0} \tag{2}$$

多项式 q 的次数比 p 的次数少1．由此式，我们有

$$p(z)=(z-z_0)q(z)+p(z_0) \tag{3}$$

把未知多项式 q 表示为

$$q(z)=b_0+b_1z+\cdots+b_{n-1}z^{n-1}$$

当把这个形式的 $q(z)$ 和类似形式的 $p(z)$ 代入到(3)式中时，可规定等式两边 z 的相同幂的系数彼此相等．由此产生了下列等式：

$$\begin{aligned} b_{n-1}&=a_n\\ b_{n-2}&=a_{n-1}+z_0b_{n-1}\\ &\vdots\\ b_0&=a_1+z_0b_1\\ p(z_0)&=a_0+z_0b_0 \end{aligned}$$

霍纳算法可写成下列紧凑形式：

```
input n, (a_i: 0≤i≤n), z_0
b_{n-1} ← a_n
for k=n-1 to 0 step -1 do
    b_{k-1} ← a_k + z_0 b_k
end do
output (b_i: -1≤i≤n-1)
```

注意在这个伪代码中 $b_{-1}=p(z_0)$．如果用铅笔和纸来完成霍纳算法中的计算，那么常常使用下面的安排．

	a_n	a_{n-1}	a_{n-2}	$\cdots$	a_0
z_0		z_0b_{n-1}	z_0b_{n-2}	$\cdots$	z_0b_0
	b_{n-1}	b_{n-2}	b_{n-3}	$\cdots$	$\boxed{b_{-1}}$

[112] 加框的数满足 $p(z_0)=b_{-1}$．

例3 用霍纳算法来求 $p(3)$，这里 p 是多项式

$$p(z)=z^4-4z^3+7z^2-5z-2$$

解 我们按上面的建议安排计算．

	1	-4	7	-5	-2
3		3	-3	12	21
	1	-1	4	7	$\boxed{19}$

因此，$p(3)=19$，并且可记

$$p(z)=(z-3)(z^3-z^2+4z+7)+19 \quad ■$$

霍纳算法也用于**降阶**．这是从多项式中除去一个线性因式的过程．若 z_0 是多项式 p 的一个根，则 $z-z_0$ 是 p 的一个因式，反之也真．p 的其他根是 $p(x)/(z-z_0)$的 $n-1$ 个根．

例 4 利用 2 是例 3 中多项式 p 的一个根这个事实，来降低多项式 p 的次数．

解 我们使用与前面的解释相同的计算排列：

	1	−4	7	−5	−2
2		2	−4	6	2
	1	−2	3	1	[0]

因此，我们有

$$z^4-4z^3+7z^2-5z-2=(z-2)(z^3-2z^2+3z+1) \quad ■$$

霍纳算法的第 3 个应用是求多项式在任意点附近的泰勒展开．设 $p(z)$是(1)式中的多项式，并且我们希望得到多项式

$$\begin{aligned}p(z)&=a_nz^n+a_{n-1}z^{n-1}+\cdots+a_0\\&=c_n(z-z_0)^n+c_{n-1}(z-z_0)^{n-1}+\cdots+c_0\end{aligned}$$

中的系数 c_k．当然，泰勒定理断言 $c_k=p^{(k)}(z_0)/k!$，但是我们要寻找效率更高的算法．注意到 $p(z_0)=c_0$，因此，通过把霍纳算法应用到多项式 p 和点 z_0 上，我们就能得到这个系数了．这算法还产生多项式

$$q(z)=\frac{p(z)-p(z_0)}{z-z_0}=c_n(z-z_0)^{n-1}+c_{n-1}(z-z_0)^{n-2}+\cdots+c_1$$

因为 $c_1=q(z_0)$，这表明第 2 个系数 c_1 可通过把霍纳算法应用到多项式 q 和点 z_0 来获得．（注意第一次应用霍纳算法产生的 q 并不是所示的形式而是 z 的幂的和形式．）重复这个过程，直到求出所有的系数 c_k． [113]

例 5 求例 3 中的多项式在点 $z_0=3$ 附近的泰勒展开．

解 这项工作可安排如下：

	1	−4	7	−5	−2
3		3	−3	12	21
	1	−1	4	7	[19]
3		3	6	30	
	1	2	10	[37]	
3		3	15		
	1	5	[25]		
3		3			
	[1]	[8]			

加框的数用在下列多项式中：

$$p(z)=(z-3)^4+8(z-3)^3+25(z-3)^2+37(z-3)+19$$ ■

我们称刚才所述的算法为**完全霍纳算法**. 改写执行它的伪代码使得系数 c_k 覆盖输入系数 a_k.

```
input n, (a_i: 0≤i≤n), z_0
for k=0 to n-1 do
    for j=n-1 to k step -1 do
        a_j ← a_j + z_0 a_{j+1}
    end do
end do
output (a_i: 0≤i≤n)
```

我们必须说明如何对多项式执行牛顿迭代. 记得这个迭代法是由下式所定义的.

$$z_{k+1}=z_k-\frac{f(z_k)}{f'(z_k)}$$

如果把这个迭代法应用于多项式 p，那么可用一种结合霍纳算法的有效方法来计算 $p(z)$和 $p'(z)$. 我们先前看到若在完全霍纳算法内仅使用两步，就能得到 $c_0=p(z_0)$和 $c_1=p'(z_0)$. 这两步可结合在伪代码中. 我们也要终止覆盖输入系数，因为在迭代的后续步中将需要它们. 假设 p 是(1)式中给定的多项式并且给定 z_0，下面是一个产生 $\alpha=p(z_0)$且 $\beta=p'(z_0)$ 的伪代码.

114

```
input n, (a_i: 0≤i≤n), z_0
α ← a_n
β ← 0
for k=n-1 to 0 step -1 do
    β ← α + z_0 β
    α ← a_k + z_0 α
end do
output α, β
```

记这个伪代码为 horner(n，(a_i：$0\leqslant i\leqslant n$)，z_0，α，β)，则从 z_0 开始对给定多项式做 M 步牛顿法的伪代码应该为：

```
input n, (a_i: 0≤i≤n), z_0, M, ε
for j=1 to M do
    call horner (n, (a_i: 0≤i≤n), z_0, α, β)
    z_1 ← z_0 - α/β
    output α, β, z_1
    if |z_1 - z_0| < ε stop
    z_0 ← z_1
end do
```

例 6 对例 3 中所用的多项式执行牛顿迭代，初始点 $z_0=0$.

解 首先，我们使用 $z_0=0$，并且用前面说明的算法计算值 $p(0)=-2$ 与 $p'(0)=-5$. z 的新值是

$$z_1=z_0-\frac{p(z_0)}{p'(z_0)}=0-\frac{-2}{-5}=-0.4$$

在具有 Marc-32 精度的计算机上执行这个算法产生了下列结果：

k	$p(z_k)$	$p'(z_k)$	z_k
1	−2.000 00	−5.000 00	−0.400 00
2	1.401 60	−12.776 00	−0.290 29
3	1.463 22	−10.173 22	−0.275 91
4	0.002 26	−9.860 30	−0.275 68
5	0.000 00	−9.855 37	−0.275 68

观察到 z_k 快速地收敛到根−0.275 68. ■

霍纳算法是正确的一个正式的证明如下. [115]

定理 5(霍纳方法定理) 设 $p(x)=a_nx^n+\cdots+a_1x+a_0$. 由算法

$$\begin{cases}(\alpha_n,\beta_n)=(a_n,0)\\(\alpha_j,\beta_j)=(a_j+x\alpha_{j+1},\alpha_{j+1}+x\beta_{j+1}) \qquad (n-1\geqslant j\geqslant 0)\end{cases}$$

定义数对$(\alpha_j,\ \beta_j)$，$j=n,\ n-1,\ \cdots,\ 0$. 则 $\alpha_0=p(x)$且 $\beta_0=p'(x)$.

证明 若 $n=0$，则 $p(x)=\alpha_0$ 并且

$$(\alpha_0,\beta_0)=(a_0,0)$$

显然有 $\alpha_0=p(x)$且 $\beta_0=p'(x)$. 因此，对 $n=0$，定理成立. 若 $n=1$，则 $p(x)=a_1x+a_0$ 并且

$$(\alpha_0,\beta_0)=(a_0+x\alpha_1,\alpha_1+x\beta_1)=(a_0+xa_1,a_1)=(p(x),p'(x))$$

因此，对 $n=1$，定理也成立. 假设对所有小于 m 的指标 n，定理成立. 设

$$p(x)=c_mx^m+\cdots+c_1x+c_0=c_0+x(c_mx^{m-1}+\cdots+c_2x+c_1)=c_0+xq(x)$$

从而，我们有

$$p'(x)=xq'(x)+q(x)$$

令 $n=m-1$，$a_n=c_m$，$a_{n-1}=c_{m-1}$，$\cdots$，$a_2=c_3$，$a_1=c_2$，$a_0=c_1$. 因为对 q，定理成立，所以我们把这算法应用于$q(x)=a_nx^n+\cdots+a_1x+a_0$. 由归纳假设得到$\alpha_0=q(x)$且$\beta_0=q'(x)$. 如果再对 p 应用这个算法，我们得到同样一组数对$(\alpha_n,\ \beta_n)$，$\cdots$，$(\alpha_1,\ \beta_1)$，$(\alpha_0,\ \beta_0)$（因为$(c_m,\ \cdots,\ c_2,\ c_1)=(a_n,\ \cdots,\ a_1,\ a_0)$)，但是在 p 的最后还要多计算一对，即

$$\begin{aligned}(\alpha_{-1},\beta_{-1})&=(a_{-1}+x\alpha_0,\alpha_0+x\beta_0)\\&=(c_0+xq(x),q(x)+xq'(x))\\&=(p(x),p'(x))\end{aligned}$$

■

注意我们可用

$$\begin{cases}(\alpha_{n-1},\beta_{n-1})=(a_{n-1}+xa_n,a_n)\\\qquad(\alpha_j,\beta_j)=(a_j+x\alpha_{j+1},\alpha_{j+1}+x\beta_{j+1})(n-2\geqslant j\geqslant 0)\end{cases}$$

或用

$$\begin{cases}(\alpha_n,\beta_n)=(a_{n-1}+xa_n,a_n)\\(\alpha_j,\beta_j)=(a_{j-1}+x\alpha_{j+1},\alpha_{j+1}+x\beta_{j+1}) \qquad (n-1\geqslant j\geqslant 1)\end{cases}$$

来定义我们的序列. 因为它们在 $n=0$ 时不做计算，所以这些定义是差的. [116]

定理6(逐次牛顿迭代定理) 设x_k和x_{k+1}是牛顿法应用于n次多项式时两个相继迭代. 则存在p的一个根在复平面内与x_k的距离小于$n|x_k-x_{k+1}|$.

证明 设$r_1,r_2,\cdots,r_n$是p的根. 于是,$p(z)=c\prod_{j=1}^{n}(z-r_j)$. 牛顿迭代中的修正项是$-p(z)/p'(z)$. p的导数是

$$p'(z)=c\sum_{k=1}^{n}\prod_{\substack{i=1\\i\neq k}}^{n}(z-r_i)=\sum_{k=1}^{n}p(z)/(z-r_k)=p(z)\sum_{k=1}^{n}(z-r_k)^{-1}$$

我们希望证明对任何z(起着x_k的作用),存在指标j使得$|z-r_j|\leqslant n|p(z)/p'(z)|$. 若没有指标$j$满足所希望的不等式,则对所有$j$,$|z-r_j|>n|p(z)/p'(z)|$. 由此可得

$$|z-r_j|^{-1}<\frac{1}{n}|p'(z)/p(z)|=\frac{1}{n}\left|\sum_{k=1}^{n}(z-r_k)^{-1}\right|\leqslant\frac{1}{n}\sum_{k=1}^{n}|z-r_k|^{-1}$$

但是这是不可能的,因为n个数的平均数不可能大于它们中的每个数. ■

这个定理归功于Bodewig[1946].

3.5.2 贝尔斯托法

即使一个多项式仅有实系数,它的根仍然可能是复数. 在这种情况下,仅使用实运算就有可能每次计算两个复根. 这个过程称为**贝尔斯托法**,接下来,将考虑这个方法.

对任何复数$z=x+iy$,其**共轭**数是$\bar{z}=x-iy$. 下面是一个将要利用的基本性质.

定理7(实二次因式定理) 若p是系数均为实数的多项式,并且w是p的一个非实数根,则$\bar{w}$也是p的一个根,并且$(z-w)(z-\bar{w})$是p的一个实二次因式.

证明 设$p(z)=a_nz^n+a_{n-1}z^{n-1}\cdots+a_1z+a_0$,其中所有的$a_k$都是实数. 因为$w$是$p$的一个根,所以我们有

$$0=a_nw^n+a_{n-1}w^{n-1}+\cdots+a_1w+a_0$$

等式两边取共轭,并且反复利用法则$\overline{z_1+z_2}=\overline{z_1}+\overline{z_2}$和$\overline{z_1z_2}=\overline{z_1}\ \overline{z_2}$. 因为$a_k$是实数,所以结果是

$$0=a_n\bar{w}^n+a_{n-1}\bar{w}^{n-1}+\cdots+a_1\bar{w}+a_0$$

因此,$\bar{w}$是p的一个根. 因为w不是实数,所以w与$\bar{w}$是p的不同的根. 从而,p含有二次因式

[117]

$$(z-w)(z-\bar{w})=z^2-(w+\bar{w})z+w\bar{w}$$

因为$w+\bar{w}$和$w\bar{w}$是实的,所以这个因式是实的. ■

实多项式p的非实根是以共轭对形式出现的,并且这些共轭对导出p的二次实因式. 因而,寻找二次因式是有道理的,这可以借助于牛顿迭代来完成.

定理8(商与余式定理) 若多项式$p(z)=a_nz^n+a_{n-1}z^{n-1}\cdots+a_1z+a_0$被二次多项式$z^2-uz-v$除,则通过令$b_{n+1}=b_{n+2}=0$然后用

$$b_k=a_k+ub_{k+1}+vb_{k+2}\qquad(n\geqslant k\geqslant 0)$$

可递归地计算商和余式

$$q(z)=b_nz^{n-2}+b_{n-1}z^{n-3}+\cdots+b_3z+b_2$$

$$r(z)=b_1(z-u)+b_0$$

证明 用

$$p(z)=q(z)(z^2-uz-v)+r(z)$$

表示 p，q 与 r 之间的关系. 这个等式详细地说明

$$\sum_{k=0}^{n}a_kz^k=\Big(\sum_{k=2}^{n}b_kz^{k-2}\Big)(z^2-uz-v)+b_1(z-u)+b_0$$

如果我们使这个等式两边 z^k 的系数相等，那么结果就是

$$\begin{aligned}a_k&=b_k-ub_{k+1}-vb_{k+2}\qquad(0\leqslant k\leqslant n-2)\\a_{n-1}&=b_{n-1}-ub_n\\a_n&=b_n\end{aligned}$$

如果我们定义 b_{n+1}和 b_{n+2}为 0，那么这三个等式中的第一个就会包括后面两个. ■

让我们专门来分析 p 中所有的系数 a_k 都是实的情况. 我们将寻找定理中提到的实二次因式的类型，因而 u 和 v 将是实的. 在上面的除法过程中，b_0 和 b_1 都是 u 和 v 的函数，并且我们记 $b_0=b_0(u,v)$和 $b_1=b_1(u,v)$. 为使 q 是 p 的一个因式，余式 r 应该为零，这导致两个方程

$$\begin{aligned}b_0(u,v)&=0\\b_1(u,v)&=0\end{aligned}$$

118

在贝尔斯托法中，用牛顿法解这对联立非线性方程. 我们需要偏导数

$$c_k=\frac{\partial b_k}{\partial u}\qquad d_k=\frac{\partial b_{k-1}}{\partial v}\qquad(0\leqslant k\leqslant n)$$

通过对定理 8 中 b_k 已经建立的递归关系微分得到上式. 这个结果就是下列一对追加的递归式：

$$\begin{aligned}c_k&=b_{k+1}+uc_{k+1}+vc_{k+2}\qquad(c_{n+1}=c_n=0)\\d_k&=b_{k+1}+ud_{k+1}+vd_{k+2}\qquad(d_{n+1}=d_n=0)\end{aligned}\tag{4}$$

因为这些递归关系显然生成同样的两个序列，所以仅需要第一个. 这个过程的概况如下：指定初值 u 和 v. 我们寻找用 δu 和 δv 表示的修正，使等式

$$b_0(u+\delta u,v+\delta v)=b_1(u+\delta u,v+\delta v)=0$$

成立. 与 3.2 节中一样，我们线性化这些等式，写成

$$b_0(u,v)+\frac{\partial b_0}{\partial u}\delta u+\frac{\partial b_0}{\partial v}\delta v=0$$

$$b_1(u,v)+\frac{\partial b_1}{\partial u}\delta u+\frac{\partial b_1}{\partial v}\delta v=0$$

鉴于前面的注释，这个方程组变成

$$\begin{bmatrix}c_0&c_1\\c_1&c_2\end{bmatrix}\begin{bmatrix}\delta u\\\delta v\end{bmatrix}=-\begin{bmatrix}b_0(u,v)\\b_1(u,v)\end{bmatrix}$$

这个方程组的解就是

$$\begin{aligned}\delta u&=(c_1b_1-c_2b_0)/J\\\delta v&=(c_1b_0-c_0b_1)/J\end{aligned}$$

$$J = c_0 c_2 - c_1^2$$

注意 J 是一对非线性方程 $b_0=b_0(u, v)$ 和 $b_1=b_1(u, v)$ 的雅可比行列式.

下面紧接着给出的是以指定点 (u, v) 为初始点，执行 M 步贝尔斯托法的伪代码.

```
input n, (a_i: 0≤i≤n), u, v, M
b_n ← a_n
c_n ← 0
c_{n-1} ← a_n
for j=1 to M do
    b_{n-1} ← a_{n-1} + u b_n
    for k=n-2 to 0 step -1 do
        b_k ← a_k + u b_{k+1} + v b_{k+2}
        c_k ← b_{k+1} + u c_{k+1} + v c_{k+2}
    end do
    J ← c_0 c_2 - c_1^2
    u ← u + (c_1 b_1 - c_2 b_0)/J
    v ← v + (c_1 b_0 - c_0 b_1)/J
    output j, u, v, b_0, b_1
end do
```

例 7 以 $(u, v)=(3, -4)$ 为初值，使用贝尔斯托法求前例中的多项式的实二次因式.

解 以双精度形式编写基于上面伪代码的计算机程序，并且在第 7 次迭代，得到

$$u = 2.275\,682\,203\,651\,0$$

$$v = -3.627\,365\,084\,711\,8$$

$$b_0 = -0.2 \times 10^{-14}$$

$$b_1 = 0.0$$

因为 b_0 和 b_1 实际上是 0，所以我们认可 u 和 v 作为二次因式 z^2-uz-v 中系数的近似值. 把这个结果与前例结合起来，就能提供多项式 p 的因式分解.（我们在等式中仅给出 3 位精度.）

$$z^4 - 4z^3 + 7z^2 - 5z - 2 = (z-2)(z+0.276)(z^2 - 2.28z + 3.63)$$

先前给出的系数值具有较高的精度. p 的两个复根可从二次因式求得(利用它的更高精度的系数)；它们是

$$1.137\,841\,101\,825\,5 \pm (1.527\,312\,250\,886\,6)\mathrm{i}$$ ■

为完成贝尔斯托法的分析，我们需要建立(在合理的假设下)雅可比行列式 J 在解点不为 0.

定理 9(贝尔斯托法中的雅可比行列式定理) 设 (u_0, v_0) 是一个使得 $z^2-u_0z-v_0$ 的根为 p 的单根的点. 则贝尔斯托法中的雅可比行列式在 (u_0, v_0) 不为 0.

119 ~ 120

证明 在这个过程中的每一步，我们有

$$p(z) = (z^2 - uz - v)q(z) + b_1(z-u) + b_0$$

关于 u 和 v 的偏微分给出下列两个方程：

$$0 = -zq(z) + (z^2 - uz - v)\frac{\partial q}{\partial u} - b_1 + \frac{\partial b_1}{\partial u}(z-u) + \frac{\partial b_0}{\partial u}$$

$$0 = -q(z) + (z^2 - uz - v)\frac{\partial q}{\partial v} + \frac{\partial b_1}{\partial v}(z-u) + \frac{\partial b_0}{\partial v}$$

采用假设条件，我们知道 $z^2-u_0z-v_0$ 有两个根，比如说 z_1 和 z_2. 由此可得 $p(z_1)=0$，$p(z_2)=0, b_1=0$，$b_0=0$. 在上面的等式中，设 $u=u_0$，$v=v_0$，且 $z=z_1$ 或 $z=z_2$. 由此产生 4 个方程

$$0=-z_jq(z_j)+c_1(z_j-u_0)+c_0 \quad (j=1,2)$$
$$0=-q(z_j)+c_2(z_j-u_0)+c_1 \quad (j=1,2)$$

其中我们直接使用了定理 8 证明中的记号和分析. 这些等式的矩阵形式是

$$\begin{bmatrix} c_0 & c_1 \\ c_1 & c_2 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ z_1-u_0 & z_2-u_0 \end{bmatrix}=\begin{bmatrix} z_1q(z_1) & z_2q(z_2) \\ q(z_1) & q(z_2) \end{bmatrix}$$

为了看出雅可比矩阵不是奇异的，只需证明上式右边的矩阵不是奇异的即可. 这个矩阵的行列式是$(z_1-z_2)q(z_1)q(z_2)$. 因为 z_1 和 z_2 是 p 的单根，所以它们不可能是 q 的根. 此外，$z_1\neq z_2$，因此，这个行列式不为 0. ■

贝尔斯托法的讨论仿效了 Henrici[1964]的说明.

3.5.3 拉盖尔迭代

我们接下来转向求多项式 p 的根的拉盖尔法. 这个方法有令人非常满意的收敛性质并且相当稳健，所以被一些现代软件包使用. 它在每个单根附近显示三阶收敛性. 这个算法是迭代的并且通过计算

$$A=-p'(z)/p(z)$$
$$B=A^2-p''(z)/p(z)$$
$$C=n^{-1}[A\pm\sqrt{(n-1)(nB-A^2)}]$$
$$z_{\text{new}}=z+1/C$$

从一个近似根 z 得到一个新的近似根. 其中 p 是一个 n 次多项式. 在 C 的定义中，选择正负号使得 $|C|$ 尽可能大.

对这个算法，下列 Kahan[1967]的定理类似于定理 6.

定理 10(拉盖尔法中的收敛半径) *若 p 是 n 次多项式，z 是任意一个复数，并且 C 按照拉盖尔算法来计算，则 p 在复平面内有一个与 z 的距离小于$\sqrt{n}/C$ 的根.* [121]

证明 像定理 6 中的逐次牛顿迭代那样，我们用 $r_1,r_2,\cdots,r_n$ 表示多项式的根. 因为 A,B,C 的公式不受附加于 p 之上的纯量因子影响，所以我们可假设 $p(x)=\prod_{j=1}^{n}(x-r_j)$，其首项系数为 1. 对 p 求导，并且令 $u_j=(r_j-z)^{-1}$，得到

$$p'(z)=\sum_{j=1}^{n}\prod_{\substack{k=1\\k\neq j}}^{n}(z-r_k)=\sum_{j=1}^{n}p(z)/(z-r_j)=-p(z)\sum_{j=1}^{n}u_j$$

因此，我们有

$$A=-\frac{p'(z)}{p(z)}=\sum_{j=1}^{n}u_j$$

对这个公式求导，并且用 B 的定义导致

$$B = \frac{-p(z)p''(z) + [p'(z)]^2}{[p(z)]^2} = \sum_{j=1}^{n}(r_j - z)^{-2} = \sum_{j=1}^{n} u_j^2$$

从 C 的定义，我们有

$$(nC - A)^2 = (n-1)(nB - A^2) \tag{5}$$

定义 $D=(A-C)/(n-1)$，因此

$$A = (n-1)D + C \tag{6}$$

从(5)式，立即可得

$$n^2C^2 - 2nCA + A^2 = (n-1)nB - nA^2 + A^2$$

$$(n-1)B = nC^2 - 2CA + A^2 \tag{7}$$

在(7)式中，用$(n-1)D+C$代替 A，得到

$$(n-1)B = nC^2 - 2C[(n-1)D + C] + [(n-1)D + C]^2$$

$$(n-1)B = nC^2 - 2(n-1)CD - 2C^2 + (n-1)^2D^2 + 2(n-1)CD + C^2$$

$$(n-1)B = (n-1)C^2 + (n-1)^2D^2$$

$$B = C^2 + (n-1)D^2 \tag{8}$$

首先用(7)，再用(6)和(8)得到

$$\begin{aligned} nB - A^2 &= nC^2 - 2CA + B \\ &= nC^2 - 2C[(n-1)D + C] + C^2 + (n-1)D^2 \\ &= (n-1)C^2 - (n-1)2CD + (n-1)D^2 \\ &= (n-1)(C-D)^2 \end{aligned} \tag{9}$$

先用(9)式，再用 A 和 B 的公式，得到

[122]

$$\begin{aligned} (n-1)|C-D|^2 &= |nB - A^2| \\ &= n^{-1}|n^2B - 2nA^2 + nA^2| \\ &= n^{-1}\left|\sum_{j=1}^{n}(n^2u_j^2 - 2nAu_j + A^2)\right| \\ &= n^{-1}\left|\sum_{j=1}^{n}(nu_j - A)^2\right| \\ &\leqslant n^{-1}\sum_{j=1}^{n}|nu_j - A|^2 \\ &= n^{-1}\sum_{j=1}^{n}\{n^2|u_j|^2 - 2n\mathrm{Re}(\overline{A}u_j) + |A|^2\} \\ &= n\sum_{j=1}^{n}|u_j|^2 - 2\overline{A}A + |A|^2 \\ &\leqslant n^2\max_j|u_j|^2 - |A|^2 \\ &= n^2\max_j|u_j|^2 - |(n-1)D + C|^2 \\ &= n^2\max_j|u_j|^2 - |(C-D) + nD|^2 \\ &= n^2\max_j|u_j|^2 - |C-D|^2 - 2n\mathrm{Re}[\overline{D}(C-D)] - n^2|D|^2 \end{aligned}$$

因此，我们有

$$n\mid C-D\mid^2\leqslant n^2\max_j\mid u_j\mid^2-n^2\mid D\mid^2-2n\mathrm{Re}(\bar{D}C)+2n\mid D\mid^2$$

$$\mid C-D\mid^2\leqslant n\max_j\mid u_j\mid^2-n\mid D\mid^2-2\mathrm{Re}(\bar{D}C)+2\mid D\mid^2$$

$$\mid C\mid^2-2\mathrm{Re}(C\bar{D})+\mid D\mid^2\leqslant n\max_j\mid u_j\mid^2-n\mid D\mid^2-2\mathrm{Re}(\bar{D}C)+2\mid D\mid^2$$

$$\mid C\mid^2+(n-1)\mid D\mid^2\leqslant n\max_j\mid u_j\mid^2=n/\min_j\mid z-r_j\mid^2$$

这给出

$$\min_j\mid z-r_j\mid^2\leqslant n/\{\mid C\mid^2+(n-1)\mid D\mid^2\}$$

以及

$$\min_j\mid z-r_j\mid\leqslant\sqrt{n}/\sqrt{\mid C\mid^2+(n-1)\mid D\mid^2}\tag{10}$$

上面这个不等式比该定理所述的不等式更强些．这个证明来自 Kahan[1967]． ■ [123]

下面提供一个指定初始点 z_0 对已知多项式执行拉盖尔法的简单的伪代码：

```
input n, (a_i: 0≤i≤n), z0, M, ε
for k=1 to M do
    α←a_n
    β←0
    γ←0
    for j=n-1 to 0 step -1 do
        γ←z0γ+β
        β←z0β+α
        α←z0α+a_j
    end do
    A←-β/α
    B←A²-2γ/α
    C←[A ± √((n-1)(nB-A²))]/n
    z1←z0+1/C
    output k, z
    if |z1-z0|<ε then stop
    z0←z1
end do
```

在这个算法中，α 是 $p(z_0)$，β 是 $p'(z_0)$，而 γ 是 $p''(z_0)/2$．

引理 1（区间端点的第一定理）　设 $v_1,v_2,\cdots,v_n$ 是任意的实数．令 $\alpha=\sum_{i=1}^{n}v_i$ 且 $\beta=\sum_{i=1}^{n}v_i^2$．则数 v_i 位于端点为

$$n^{-1}\left[\alpha\pm\sqrt{(n-1)(n\beta-\alpha^2)}\right]\tag{11}$$

的闭区间内．

证明　只需证明 v_1 位于所述的区间内即可．回忆柯西-施瓦茨不等式：

$$\Big(\sum_{i=1}^{m}x_iy_i\Big)^2\leqslant\Big(\sum_{i=1}^{m}x_i^2\Big)\Big(\sum_{j=1}^{m}y_j^2\Big)$$

运用它，我们有

$$
\begin{aligned}
\alpha^2 - 2\alpha v_1 + v_1^2 &= (\alpha - v_1)^2 = (v_2 + v_3 + \cdots + v_n)^2 \\
&\leqslant (1^2 + 1^2 + \cdots + 1^2)(v_2^2 + v_3^2 + \cdots + v_n^2) \\
&= (n-1)(v_2^2 + v_3^2 + \cdots + v_n^2) \\
&= (n-1)(\beta - v_1^2) - (n-1)\beta - nv_1^2 + v_1^2
\end{aligned}
$$

整理这个不等式，得

124

$$nv_1^2 - 2\alpha v_1 + \alpha^2 - (n-1)\beta \leqslant 0$$

这说明二次函数 $q(x)=nx^2-2\alpha x+\alpha^2-(n-1)\beta$ 有性质 $q(v_1)\leqslant 0$. 对大的 $|x|$，显然有 $q(x)>0$. 因此，v_1 位于 q 的两个根之间，并且它们是公式(11)中的端点. ■

引理 2(区间端点的第二定理) *设 p 是 n 次实多项式，其根 $r_1, r_2, \cdots, r_n$ 是实的. 对任何与所有 r_j 不同的实数 x，则数 $(x-r_j)^{-1}$ 位于端点为*

$$[np(x)]^{-1}\left\{p'(x) \pm \sqrt{[(n-1)p'(x)]^2 - n(n-1)p(x)p''(x)}\right\} \tag{12}$$

的区间内.

证明 多项式 p 具有 $p(x)=c\prod_{j=1}^{n}(x-r_j)$ 形式. 因为要证明的论断与 c 无关，所以假设 $c=1$. 像定理 10 中关于拉盖尔法收敛半径的证明一样，用 $v_j=(x-r_j)^{-1}$，$\alpha=\sum_{j=1}^{n}v_j$ 且 $\beta=\sum_{j=1}^{n}v_j^2$，我们有

$$p'(x)/p(x) = \sum_{j=1}^{n} v_j = \alpha \tag{13}$$

$$[(p'(x))^2 - p(x)p''(x)]/(p(x))^2 = \sum_{j=1}^{n} v_j^2 = \beta \tag{14}$$

根据引理 1，所有的数 v_j 都位于以公式(11)给出的数为端点的区间内. 当把(13)式和(14)式中的值 α 和 β 代入这个公式时，结果就是公式(12). ■

定理 11(拉盖尔方法的单调收敛性定理) *设 p 是实多项式，其根都是实的. 由拉盖尔算法以任意初始点产生的序列单调地收敛于 p 的根.*

证明 设 p 的根为 $r_1\leqslant r_2\leqslant\cdots\leqslant r_n$. 假设 x 不是根. 则根据引理 2，所有的数 $(x-r_i)^{-1}$ 位于端点为

$$u(x) = [p'(x) + w(x)]/np(x) \text{ 和 } v(x) = [p'(x) - w(x)]/np(x)$$

的区间内，其中

$$w(x) = \sqrt{[(n-1)p'(x)]^2 - n(n-1)p''(x)p(x)}$$

首先，让我们考虑对某个 j，$r_j<x<r_{j+1}$ 的情况：假设在区间 (r_j, r_{j+1}) 上，$p(y)>0$，其他情况是类似的. 于是有

125

$$v(x) \leqslant (x - r_{j+1})^{-1} < 0 < (x - r_j)^{-1} \leqslant u(x)$$

由此立即可得

$$r_j \leqslant x - \frac{1}{u(x)} < x < x - \frac{1}{v(x)} \leqslant r_{j+1}$$

因而，若我们从 (r_j, r_{j+1})内的 x 开始，并且按照拉盖尔迭代计算 $x-1/u(x)$和 $x-1/v(x)$，则这两个新点将位于$[r_j, r_{j+1}]$内．当然，如果这两个新点中任一个是这个区间的一个端点，那么 p 的根就已经算出．反之，这两个新点位于(r_j, r_{j+1})内，可以重复这个过程．形式上依次进行，我们令

$$\begin{cases} y_0 = x \\ y_{k+1} = y_k - 1/u(y_k) \qquad (k \geqslant 1) \end{cases}$$

和

$$\begin{cases} z_0 = x \\ z_{k+1} = z_k - 1/v(z_k) \qquad (k \geqslant 1) \end{cases}$$

来定义两个序列$[y_k]$和$[z_k]$．上面的分析说明

$$r_j \leqslant y_{k+1} < y_k < \cdots < x < z_1 < \cdots < z_k < z_{k+1} \leqslant r_{j+1}$$

因为序列$[y_k]$递减并以且 r_j 为下界，所以它必收敛．设 $y_k \downarrow y$．从迭代公式，我们有

$$1/u(y_k) = y_k - y_{k+1} \to 0$$

因此，$u(y_k)\to\infty$．现在对 u 的公式表明 $p(y_k)\to 0$，所以 $p(y)=0$ 且 $y=r_j$．类似地，我们有结论 $z_k \uparrow r_{j+1}$．

余下的情况就是初始点不在 p 的两个根之间．设 $x>r_n$．还假设在(r_n, ∞)上，$p(y)>0$．如前所述进行，有

$$0 < (x-r_n)^{-1} \leqslant u(x)$$

由此立即可得 $r_n \leqslant x-1/u(x)<x$．同理，如前所定义的序列$[y_k]$向下收敛于 r_n．■

在 Bodewig[1946]、van der Corput[1946]、Durand[1960]、Foster[1981]、Galeone[1977]、Householder[1970]、Kahan[1967]、Ostrowski[1966]、Parlett[1964]、Redish[1974]以及 Wilkinson[1965]等论文和著作中可找到有关拉盖尔法的讨论．

3.5.4 复牛顿法

对复系数的多项式，牛顿法应该以复数运算来编程．在求出一个根之后，应该使用降阶过程(也是以复数运算编程)．因此牛顿法能应用于降阶后的多项式．这个过程可以重复直到确定所有的根为止．进一步分析和实验指出，一般说来，如果采取了下面两个措施，那么这个过程是相当令人满意的：

126

1. 按值的递增次序来计算根．

2. 用根的最佳估计作为初值，通过对原始多项式应用牛顿法来直接改进在降阶后的多项式上使用牛顿法所得到的任何根．并只在完成这些以后，才执行下一步降阶．

关于预防措施总体方法的进一步说明，我们推荐读者参考 Wilkinson[1984]和 Peters and Wilkinson[1971]．对于多项式求根问题，可以讨论其他许多方法．拉盖尔法是特别吸引人的，因为它有良好的整体收敛性态．Jenkins and Traub[1970a]已经推出了一种适合通用软件的稳健方法．其他没有提到的文献有 Allgower，Glasshoff，and Peitgen[1981]、Gautschi[1979，1984]、Henrici[1974]、Householder[1970]、Ostrowski[1966]、Jenkins and Traub[1970b]、Marden[1966]、Smale[1981]、Stoer and Bulirsch[1980]、Ralston and Rabinowitz[1978]以及 Traub[1964]．

这里有一个怎样得到图 3-8 的简短解释. 它涉及复平面内的牛顿法.

设 p 是一个至少 2 次的多项式，并设 ξ 是它的一个根. 如果牛顿法从复平面内的一点 z 开始，那么它就产生一个由等式

$$\begin{cases} z_0 &= z \\ z_{n+1} &= z_n - p(z_n)/p'(z_n) \qquad (n \geqslant 0) \end{cases}$$

定义的序列. 如果 $\lim_{n\to\infty} z_n = \xi$，我们说 z(初始点)被**吸引**到 ξ. 所有被吸引到 ξ 的点的集合称为对应于 ξ 的**吸引盆**. p 的每个根都有一个吸引盆，并且它们是彼此不相交的集合，因为收敛于 p 的一个根的序列不会再收敛于其另一根. 有些复数不属于任何吸引盆，它们就是使得牛顿法不收敛的初始点. 这些例外点构成 p 的**茹利亚集**. 之所以这样命名是为了纪念法国数学家 G. Julia(茹利亚). 她于 1918 年发表了一本关于这个主题的重要论文集. 若 p 的所有根都是单的，则吸引盆是复平面内的开集，并且茹利亚集是每个吸引盆的边界.

127

多项式 $p(z)=z^5-1$ 的吸引盆如图 3-8 所示. p 的 5 个根是

$$\omega_i = \cos\left(\frac{2}{5}\pi k\right) + \mathrm{i}\sin\left(\frac{2}{5}\pi k\right) \qquad k = 0,1,2,3,4$$

我们在复平面内取一个正方形区域并在这个正方形中生成大量的格点. 对于每个格点，我们进行一个粗略的试验来确定它属于哪个吸引盆. 这个试验由两部分组成：记录前 20 次牛顿迭代和检验每个迭代值与一个根的距离是否都在 0.25 之内. 果真如此的话，那么随后的迭代就二次收敛于那个根. 在复平面中，这个事实可以用牛顿法的标准收敛理论来证实. 以这种方式，生成了一列属于每个吸引盆的格点. 对根 ω_0 的吸引盆指定一种颜色，并且对根 ω_1，ω_2，ω_3，ω_4 的吸引盆分别指定其他四种颜色. 在工作站的彩屏上显示出这 5 个吸引盆(实际上仅仅是它们中的格点)，由此在屏幕上形成了一幅彩图. 这 5 个彩色集以不可思议的方式组合在一起显示了一种**分形**现象. 即放大两个集合相交处所在平面的一部分，我们看到重复着同样的通用图案. 在反复放大的过程中，这个图形依然保持不变. 此外，这 5 个集合中每一个的边界点也是其他 3 个集合的边界点！

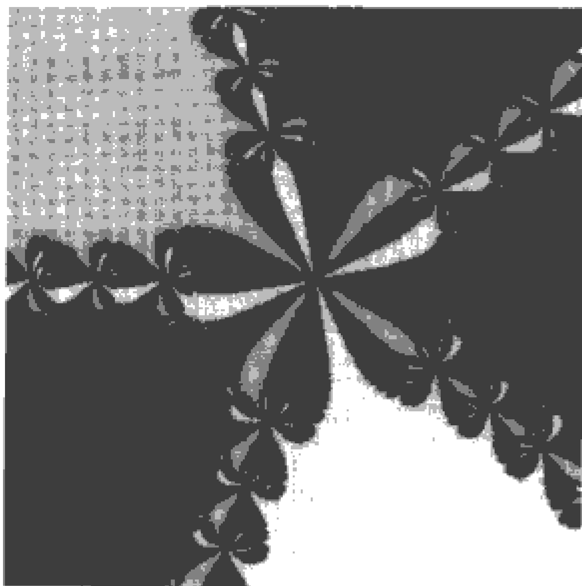


图 3-8 多项式 $p(z)=z^5-1$ 的吸引盆

最近，出版了一些关于分形和混沌的论文与书籍. 更多的资料可以在 Barnsley[1988]、Curry, Garnett, and Sullivan[1983]、Dewdney[1988]、Glieck[1987]、Mandelbrot[1982]、Peitgen and Richter[1986]、Peitgen, Saupe, and Haeseler[1984]、Pickover[1988]以及Sander[1987]中找到.

习题 3.5

1. 用霍纳算法求 $p(4)$，这里

$$p(z)=3z^5-7z^4-5z^3+z^2-8z+2$$

2. (续)对上题的多项式，求泰勒级数关于点 $z_0=4$ 的展开式.
3. (续)对(上面)习题 3.5.1 的多项式，在点 $z_0=4$ 开始牛顿法. 试问 z_1 是多少?
4. (续)对(上面)习题 3.5.1 的多项式，运用贝尔斯托法，初始点为$(u, v)=(3, 1)$. 计算修正值 δu 和 δv.
5. (续)对(上面)习题 3.5.1 的多项式，求一个中心在原点包含所有根的圆盘.
6. (续)对(上面)习题 3.5.1 的多项式，求一个中心在原点不包含任何根的圆盘.
7. 定理 3 有时是否能给出圆心在原点、包含已知多项式所有根的最小圆的半径? [128]
8. 证明每个实系数多项式能因式分解成实系数线性因式和实系数二次因式的积.
9. 验证贝尔斯托法讨论过程中的递归关系 c_k 和 d_k 以及给定的初始值.
10. 对多项式 $p(z)=9z^4-7z^3+z^2-2z+5$，计算 $p(6)$，$p'(6)$，以及从 $z=6$ 开始用牛顿迭代计算中的下一个点.
11. 用课文中所采用的多重性定义，证明若 z 是多项式 p 的 m 重根，则 $p(z)=p'(z)=\cdots=p^{(m-1)}(z)=0$ 而 $p^{(m)}(z)\neq 0$.
12. (续)证明上题的逆命题.
13. 贝尔斯托法是否产生二次收敛的序列(u_k, v_k)?
14. 当根 z_0 已知时，编写一个降低 $p(z)$次数的算法，但是以升幂方式——首先是常数项，计算降阶后的多项式的系数.
15. 推导贝尔斯托法讨论过程中的(4)式.
16. 就多项式 $p(x)=a_0+a_1x+\cdots+a_nx^n$ 而言，证明下列结论. 对给定的 x，我们令$(\alpha_n, \beta_n, \gamma_n)=(a_n, 0, 0)$并且归纳定义

$$(\alpha_j,\beta_j,\gamma_j)=(a_j+x\alpha_{j+1},\alpha_{j+1}+x\beta_{j+1},\beta_{j+1}+x\gamma_{j+1})$$

$j=n-1, n-2, \cdots, 0$. 则 $p(x)=\alpha_0$，$p'(x)=\beta_0$，$p''(x)=2\gamma_0$.

17. 证明在拉盖尔法的分析中的

$$C^2+(n-1)D^2=\sum_{j=1}^{n}u_j^2$$

18. 在拉盖尔法的描述中，量 A 和 B 都是 z 的函数并且依赖所给定的多项式 p. 设 r 是 p 的根. 证明 $p(z)/(z-r)$的对应函数 A 和 B 分别是

$$A+(z-r)^{-1} \text{和} B-(z-r)^2$$

19. 证明若 p 是一个实系数 n 次多项式，则

$$(n-1)[p'(x)]^2\geqslant np(x)p''(x)$$

这里假定所有根都是实的.

计算机习题 3.5

1. 编写一个程序，多项式 p 的系数和特殊点 z_0 作为输入，而产生的值 $p(z_0)$，$p'(z_0)$，$p''(z_0)$ 作为输出. 编写只有一次循环的伪代码. 用习题 3.5.1 中的多项式作试验，取 $z_0=4$.

2. 编写一个在复平面内给定初始点和给定迭代次数的复系数多项式的复牛顿法. 用习题 3.5.1 的多项式对你的程序进行测试，其中初始点为 $z_0=3-2i$.
3. 用按照复数运算编码的拉盖尔法进行实验. 求出本节例题中使用的多项式的所有 4 个根.
4. 使用牛顿法和多项式 $p(z)=z^3-1$，求 3 个邻近的初始点(彼此相距 0.01 以内)使得由此产生的序列收敛于不同的根. 用线段连接相继的点，在图形显示器上显示这些点的序列在一个包含根的正方形内的轨迹.

129

5. 编写一个使用牛顿法和降阶求多项式所有根的计算机程序. 用计算机习题 2.3.6 中 Wilkinson 的 Perfidious 多项式对这程序进行测试.
6. 编写并且测试一个计算复多项式所有根的程序. 使用拉盖尔法以根的模增大的次序计算根. 使用降阶，并且通过对原始多项式的拉盖尔迭代来改进从已降阶的多项式中获得的根. 为了测试你的程序，计算多项式

$$x^8-36x^7+546x^6-4\,536x^5+22\,449x^4-67\,284x^3$$
$$+118\,124x^2-109\,584x+40\,320$$

的所有根. 正确的根是整数 1，2，…，8. 接下来，当把 x^7 的系数改成 -37 时，求解同样的方程. 观察系数中较小的扰动是怎样引起根中的巨大变化的. 因此，根是系数的不稳定函数.
7. 修正拉盖尔算法使得在霍纳算法的应用中保存商多项式的系数. 在求出一个根之后，就没有必要对已降阶的多项式单独进行计算了. 用

$$z^4-(10+26i)z^3-(216-190i)z^2+(1\,140+636i)z-(72-68i)=0$$

测试你的程序.
8. 像在本节中所讨论的那样，编写一个使用复牛顿法来计算 $f(z)=z^8+1$ 根的吸引盆的计算机程序. 在一台图形计算机终端或图形工作站上显示这些结果.

3.6 同伦法和延拓法

在本节，我们考虑下列方程求根的问题

$$f(x)=0 \tag{1}$$

这里 f 是从一个线性空间到另一个线性空间的映射，比如，$f: X\to Y$. 这是一个包括代数方程组、积分方程组、微分方程组等非常一般的问题. 这里考虑的方法与前面各节中所采用的策略稍有不同，并且这个新方法还需要常微分方程组的数值解，在本书中这个主题直到 8.6 节才开始讨论. 此外，有关这个方法的一个例子涉及线性规划问题，我们将在 10.3 节中讨论这个问题.

3.6.1 基本概念

延拓法的基本思想是使用一个取遍区间[0，1]的参数 t，把一个已知问题嵌入到一个单参数的问题族中. 先安排原始问题与 $t=1$ 对应，再安排一个有已知解的问题与 $t=0$ 对应. 例如，我们可定义

130

$$h(t,x)=tf(x)+(1-t)g(x) \tag{2}$$

方程 $g(x)=0$ 应该有一个已知解. 下一步是选择点 t_0，t_1，…，t_m 使得

$$0=t_0<t_1<t_2<\cdots<t_m=1$$

然后，我们试图求解每个方程 $h(t_i, x)=0$，$1\leqslant i\leqslant m$. 假定将使用某种迭代方法(比如牛顿法)，那么用第 i 步的解作为计算第 $i+1$ 步解的初始点是明智的.

这整个过程可看作是解决困难的一个对策，牛顿法的困难就是需要好的初始点.

关系(2)把原始问题(1)嵌入到一个问题族，它是连接两个函数 f 和 g **同伦**的一个实例. 一

般而言，同伦可以是 f 和 g 之间任何连续的连接．正式地讲，两个函数 f，g：$X\to Y$ 之间的同伦是一个连续的映射

$$h:[0,1]\times X\to Y \tag{3}$$

使得 $h(0,x)=g(x)$ 且 $h(1,x)=f(x)$．如果这样的映射存在，我们就说 f 是与 g **同伦的**．这是一个从 X 到 Y 连续映射之间的等价关系，此处 X 和 Y 可以是任意两个拓扑空间．

一个常常用于延拓法的简单同伦是

$$\begin{aligned} h(t,x) &= tf(x)+(1-t)[f(x)-f(x_0)] \\ &= f(x)+(t-1)f(x_0) \end{aligned} \tag{4}$$

这里 x_0 可以是 X 中的任意点，并且显然 x_0 是 $t=0$ 时问题的一个解．

若对于每个 $t\in[0,1]$，方程 $h(t,x)=0$ 有唯一的根，则那个根就是 t 的一个函数，并且可以记 $x(t)$ 为使方程 $h(t,x(t))=0$ 成立的 X 的唯一成员．集合

$$\{x(t):0\leqslant t\leqslant 1\} \tag{5}$$

可解释为 X 中用参数 t 表示的弧或曲线．这条弧从已知点 $x(0)$ 开始到问题的解 $x(1)$ 结束．延拓法试图通过计算这条曲线上的点 $x(t_0)$，$x(t_1)$，…，$x(t_m)$ 来确定这条曲线．

若函数 $t\mapsto x(t)$ 是可微的，h 也是可微的，则隐函数定理使得我们能计算 $x'(t)$．沿着这个思路，我们可用微分方程来刻画(5)中的曲线．假定任意的同伦，我们有

$$0=h(t,x(t)) \tag{6}$$

关于 t 求导，得到

$$0=h_t(t,x(t))+h_x(t,x(t))x'(t) \tag{7}$$

其中下标表示偏导．因而，

$$x'(t)=-[h_x(t,x(t))]^{-1}h_t(t,x(t)) \tag{8}$$

131

这是一个关于 x 的微分方程．它有已知的初值，因为 $x(0)$ 是假定已知的．则对这个微分方程进行积分(通常利用数值过程)，得到值 $x(1)$，它就是解．

例 1　设 $X=Y=\mathbb{R}^2$，并且定义

$$f(x)=\begin{bmatrix}\xi_1^2-3\xi_2^2+3\\ \xi_1\xi_2+6\end{bmatrix}\qquad x=(\xi_1,\xi_2)\in\mathbb{R}^2$$

解　等式(4)给出一个方便的同伦，我们选择 $x_0=(1,1)$．计算(8)式右边的导数是

$$h_x=f'(x)=\begin{bmatrix}\partial f_1/\partial\xi_1 & \partial f_1/\partial\xi_2\\ \partial f_2/\xi_1 & \partial f_2/\xi_2\end{bmatrix}=\begin{bmatrix}2\xi_1 & -6\xi_2\\ \xi_2 & \xi_1\end{bmatrix}$$

$$h_t=f(x_0)=\begin{bmatrix}f_1(x_0)\\ f_2(x_0)\end{bmatrix}=\begin{bmatrix}1\\ 7\end{bmatrix}$$

$f'(x)$的逆是

$$h_x^{-1}=[f'(x)]^{-1}=\frac{1}{\Delta}\begin{bmatrix}\xi_1 & 6\xi_2\\ -\xi_2 & 2\xi_1\end{bmatrix}\qquad \Delta=2\xi_1^2+6\xi_2^2$$

控制引导远离点 x_0 路径的微分方程是(8)式．在此情形下，它是一对常微分方程：

$$\begin{bmatrix}\xi_1'\\ \xi_2'\end{bmatrix}=-\frac{1}{\Delta}\begin{bmatrix}\xi_1 & 6\xi_2\\ -\xi_2 & 2\xi_1\end{bmatrix}\begin{bmatrix}1\\ 7\end{bmatrix}=-\frac{1}{\Delta}\begin{bmatrix}\xi_1+42\xi_2\\ -\xi_2+14\xi_1\end{bmatrix}$$

在区间 $0\leqslant t\leqslant 1$ 上，对这个方程组数值积分(使用第 8 章所介绍的任何一种方法)，并且在 $t=1$ 处解是(−2.961，1.978). 注意 f 有根(−3，2).

我们可以从由同伦方法产生的点开始进行牛顿迭代来结束这个问题，并且求出它的数值解. 牛顿迭代用 $x-\delta$ 来代替任何近似根 x，这里给出的修正 δ 为

$$\delta=[f'(x)]^{-1}f(x)$$

在此例中，向量 δ 是

[132]

$$\begin{bmatrix}\delta_1\\ \delta_2\end{bmatrix}=\frac{1}{\Delta}\begin{bmatrix}\xi_1 & 6\xi_2\\ -\xi_2 & 2\xi_1\end{bmatrix}\begin{bmatrix}\xi_1^2-3\xi_2^2+3\\ \xi_1\xi_2+6\end{bmatrix}$$

3 步牛顿迭代产生下列结果：

k	ξ_1	ξ_2
0	−2.961 000 000 000	1.978 000 000 000
1	−3.000 253 281 314	2.000 320 274 478
2	−3.000 000 005 780	2.000 000 037 824
3	−3.000 000 000 000	2.000 000 000 000

■

下列结果属于 Ortega and Rheinboldt[1970]，在他们所给的条件下，同伦方法将会成功.

定理 1(连续可微解定理) *若 f: $\mathbb{R}^n\to\mathbb{R}^n$ 是连续可微的并且在 $\mathbb{R}^n$ 上 $\|[f'(x)]^{-1}\|\leqslant M$，则对任意 $x_0\in\mathbb{R}^n$，在 $\mathbb{R}^n$ 中存在唯一的曲线 $\{x(t)\ :\ 0\leqslant t\leqslant 1\}$ 使得 $f(x(t))+(t-1)f(x_0)=0$，其中 $0\leqslant t\leqslant 1$. 函数 $t\mapsto x(t)$ 是初值问题 $x'=-[f'(x)]^{-1}f(x_0)$ 的连续可微解，这里 $x(0)=x_0$.*

3.6.2 跟踪路径

跟踪路径 $x(t)$ 的另一种方法是由 Garcia and Zangwill[1981]给出的. 我们从方程 $h(t, x)=0$ 开始，假定 $x\in\mathbb{R}^n$，并且 $t\in[0, 1]$. 向量 $y\in\mathbb{R}^{n+1}$ 定义为

$$y=(t,\xi_1,\xi_2,\cdots,\xi_n)$$

这里 ξ_1，ξ_2，…，ξ_n 是 x 的分量. 因此，方程就是 $h(y)=0$. y 的每个分量，包括 t，允许是一个独立变量 s 的函数，并且可写成 $h(y(s))=0$. 关于 s 求导，就得到基本微分方程

$$h'(y(s))y'(s)=0 \tag{9}$$

变量 s 如 t 那样，从 0 开始，x 的初值是 $x(0)=x_0$. 因而微分方程(9)有合适的初值.

因为 f 和 g 是 $\mathbb{R}^n$ 到 $\mathbb{R}^n$ 内的映射，h 是 $\mathbb{R}^{n+1}$ 到 $\mathbb{R}^n$ 内的映射. 因此，导数 $h'(y)$ 用一个 $n\times(n+1)$ 矩阵 A 来表示. 向量 $y(s)$ 有 $n+1$ 个分量，我们用 η_1，η_2，…，η_{n+1} 来表示它们. 那么利用下面的引理，我们可得到方程(9)的另一种形式，即

$$\eta_j'=(-1)^{j+1}\det(A_j)\qquad(1\leqslant j\leqslant n+1) \tag{10}$$

这里 A_j 是把 A 的第 j 列划掉后所得到的 $n\times n$ 矩阵. 下面用例 1 中的同样问题来说明这个形式记号.

例 2 取例 1 中的 f 和 x_0，我们有

[133]

$$h(t,x)=\begin{bmatrix}\xi_1^2-3\xi_2^2+2+t\\ \xi_1\xi_2-1+7t\end{bmatrix}$$

解 由

$$\begin{bmatrix} \partial h_1/\partial t & \partial h_1/\partial \xi_1 & \partial h_1/\partial \xi_2 \\ \partial h_2/\partial t & \partial h_2/\partial \xi_1 & \partial h_2/\partial \xi_2 \end{bmatrix} \begin{bmatrix} t' \\ \xi_1' \\ \xi_2' \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2\xi_1 & -6\xi_2 \\ 7 & \xi_2 & \xi_1 \end{bmatrix} \begin{bmatrix} t' \\ \xi_1' \\ \xi_2' \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{11}$$

给出微分方程(9). 然而使用方程(10)更可取，我们把微分方程写成下列形式

$$\begin{cases} t' = 2\xi_1^2 + 6\xi_2^2 & t(0) = 0 \\ \xi_1' = -\xi_1 - 42\xi_2 & \xi_1(0) = 1 \\ \xi_2' = \xi_2 - 14\xi_1 & \xi_2(0) = 1 \end{cases} \tag{12}$$

上面方程组中的导数是关于 s 的. 执行数值积分，我们得到：

$$s = 0.087 \quad t = 0.969 \quad \xi_1 = -2.94 \quad \xi_2 = 1.97$$

$$s = 0.088 \quad t = 1.010 \quad \xi_1 = -3.02 \quad \xi_2 = 2.01$$

如例 1 中所做的那样，它们中的每一个都能用来开始进行牛顿迭代. 这个同伦方法所生成的路径如图 3-9 所示.

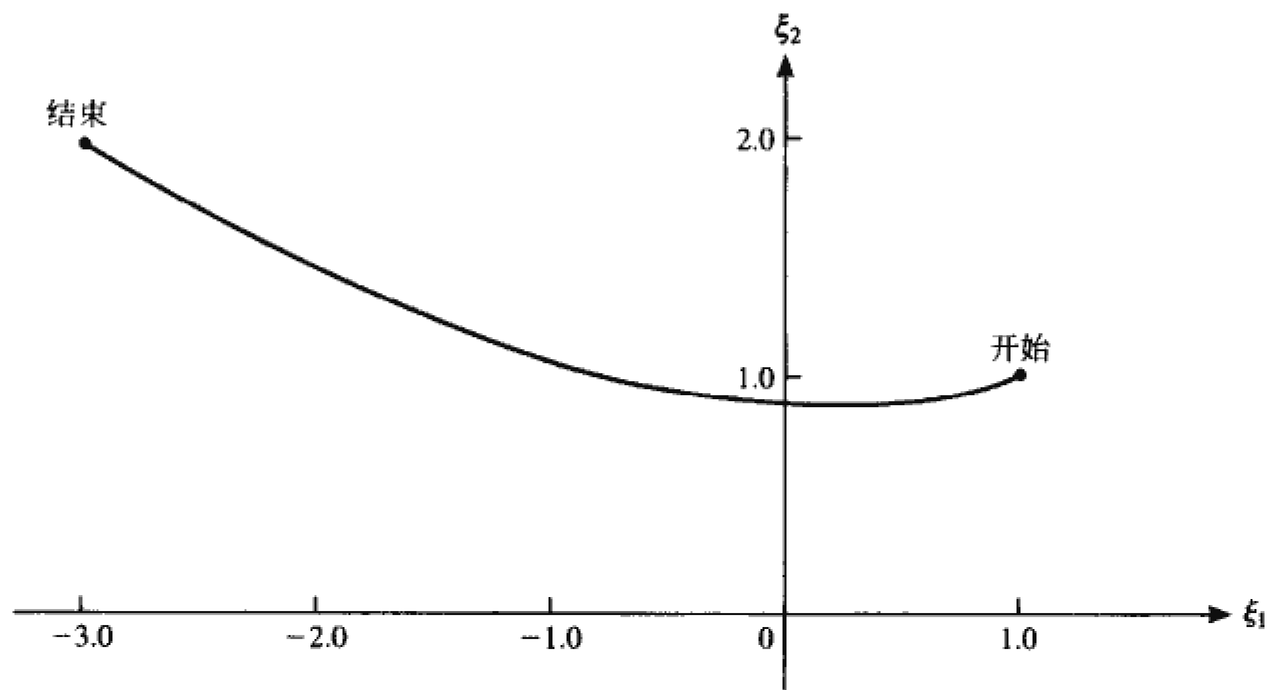


图 3-9 例 2 中生成的路径

■

使用例 2 中方法的缺点是我们对相应于 $t=1$ 的 s 值不能事先知道. 在实践中，这可能就需要好几台计算机来运行.

134

引理 1(齐次方程的解引理) 设 A 是一个 $n\times(n+1)$ 矩阵. 齐次方程 $Ax=0$ 的解是由 $x_j=(-1)^j\det(A_j)$ 给出的，这里 A_j 是 A 中去掉第 j 列后得到的矩阵.

证明 选择 A 中任意一行(比如，第 i 行)，把它作为新的一行添加在 A 的顶端. 这就产生了一个 $(n+1)\times(n+1)$ 矩阵 B，因为 A 的第 i 行在 B 中出现了两次，所以它显然是奇异的.

按 B 的第 1 行展开它的行列式，我们得到

$$0 = \det B = \sum_{j=1}^{n+1}(-1)^{j+1}a_{ij}\det(A_j) = -\sum_{j=1}^{n+1}a_{ij}x_j$$

因为这对 $i=1, 2, \cdots, n$ 都是成立的，所以我们有 $Ax=0$. ■

3.6.3 与牛顿法的关系

同伦方法与牛顿法之间的关系要比直观所见深刻的多. 让我们从同伦

$$h(t,x) = f(x) - e^{-t}f(x_0) \tag{13}$$

开始. 在此式中，t 将取遍 0 到∞的所有值. 我们寻找一条曲线或路径 $x=x(t)$，使得在其上有

$$0 = h(t,x(t)) = f(x(t)) - e^{-t}f(x_0)$$

通常，关于 t 的微分将产生一个刻画路径的微分方程：

$$\begin{aligned} 0 &= f'(x(t))x'(t) + e^{-t}f(x_0) \\ &= f'(x(t))x'(t) + f(x(t)) \\ x'(t) &= -[f'(x(t))]^{-1}f(x(t)) \end{aligned} \tag{14}$$

如果用欧拉方法(第 8 章中说明)对这个微分方程求积，其中步长取 1，那么结果就是公式

$$x_{n+1} = x_n - [f'(x_n)]^{-1}f(x_n)$$

显然这是牛顿法的公式. 当然人们可以期望用更精确的数值方法和可变的步长来对 (14)式求积以得到更好的结果.（有关这些内容，参考第 8 章.）

3.6.4 线性规划

同伦方法可用来解线性规划问题.（这样的问题在 10.3 节中讨论.）这个方法自然地导致出由 Karmarkar[1984]提出的算法. 在本文中解释同伦方法时，我们较完整地沿用 Brophy and Smith[1988]所给出的描述. 希望进一步研究这些思想的读者可以参考其他文献.

135

考虑标准线性规划问题

$$\begin{cases} 最大化\ c^{\mathrm{T}}x \\ Ax = b, x \geqslant 0 \end{cases} \tag{15}$$

这里，$c\in\mathbb{R}^n$，$x\in\mathbb{R}^n$，$b\in\mathbb{R}^m$，并且 A 是一个 $m\times n$ 矩阵. 我们从**可行点**(满足约束条件的点 x^0)开始. **可行集**是

$$\mathcal{F} = \{x \in \mathbb{R}^n : Ax = b\ 和\ x \geqslant 0\}$$

我们的意图是从 $x^{(0)}$ 移动到始终保持在 $\mathcal{F}$ 内的其他后续点，并且增加**目标函数** $c^{\mathrm{T}}x$ 的值. 显然如果从 $x^{(0)}$ 移动到 $x^{(1)}$，差 $x^{(1)}-x^{(0)}$ 必然就在 A 的零空间内. 我们将设法在可行集中找到一条曲线 $t\mapsto x(t)$，从 x^0 开始，导出极值问题的一个解. 我们要求必须具备下列条件：

1. $x(t)\geqslant 0$，$t\geqslant 0$
2. $Ax(t)=b$，$t\geqslant 0$
3. 对 $t\geqslant 0$，$c^{\mathrm{T}}x(t)$递增

这条曲线用初值问题

$$\begin{cases} x' = f(x) \\ x(0) = x^{(0)} \end{cases} \tag{16}$$

来定义．我们面临的任务是确定一个适当的 f．为满足条件 1，我们约定当分量 x_i 接近 0 时，它的速度 $x'_i(t)$ 也接近 0．这可通过令

$$D(x) = \begin{bmatrix} x_1 & & & 0 \\ & x_2 & & \\ & & \ddots & \\ 0 & & & x_n \end{bmatrix}$$

和假定对某个有界函数 G，有

$$f(x) = D(x)G(x) \tag{17}$$

来达到目的．若是这种情况，则从(16)和(17)式，我们有

$$x'_i = x_i G_i(x)$$

且在 $x_i \to 0$ 的条件下，显然 $x'_i \to 0$．

为满足必要条件 2，只需要求 $Ax' = 0$ 即可．因为 $x' = f = DG$，所以要求 $ADG = 0$．其最方便的安排是令 $G = PH$，这里 H 是任何函数而 P 是到 AD 零空间上的正交投影． 136

最后，为满足条件 3，选择 H 以便 $c^T x(t)$ 递增．因而，我们希望

$$0 < \frac{\mathrm{d}}{\mathrm{d}t}(c^T x(t)) = c^T x' = c^T f(x) = c^T DG = c^T DPH$$

Dc 是 H 的方便选择，用 $v = Dc$，我们有

$$\begin{aligned} c^T DPH &= c^T DPDc = v^T Pv = \langle v, Pv \rangle \\ &= \langle v - Pv + Pv, Pv \rangle = \langle Pv, Pv \rangle \geqslant 0 \end{aligned}$$

注意 $v - Pv$ 正交于 P 的值域，于是 $\langle v - Pv, Pv \rangle = 0$．

那么初值问题的最终形式是

$$x' = D(x)P(x)D(x)c \qquad x(0) = x^{(0)} \tag{18}$$

P 的理论上的公式是

$$P = I - (AD)^T[(AD)(AD)^T]^{-1}AD \tag{19}$$

它的正确性依赖于 $B \equiv AD$ 满秩，使得 BB^T 非奇异．同样这将要求每个分量 $x_i > 0$．因而，点 $x(t)$ 应该继续保留在集合

$$\{x : x \geqslant 0\}$$

的内部．特别是 $x^{(0)}$ 就应该这样选择．实际上，Pv 不用(19)式来计算，而是通过解方程 $BB^T z = Bv$ 和用

$$Pv = v - B^T z$$

来计算．

初值问题(18)不需要非常精确地求解．可以使用欧拉方法的一种变形．回忆针对方程(16)欧拉方法是用

$$x(t+\delta) = x(t) + \delta x'(t) = x(t) + \delta f(x)$$

来向前推导解．利用此类公式，我们可由等式

$$x^{(k+1)} = x^{(k)} + \delta_k f(x^{(k)})$$

生成向量序列 $x^{(0)}$，$x^{(1)}$，…．虽然 δ_k 的值取得越大越能达到必备的条件 $x^{(k+1)} \in \mathcal{F}$，但那将使得点 $x^{(k+1)}$ 至少要有一个 0 分量．如前所述，这又将导致其他的麻烦．在实践中最适当的方法是近似地取 δ_k 为最大可能步的 9/10．因为最大可能步是满足 $x^{(k+1)} \geqslant 0$ 的最大 λ，所以很容易计算出来．（约束 $Ax=b$ 被自动地维持．）

137

习题 3.6

1. 用例 2 中使用的同伦方法解方程组

$$x-2y+y^2+y^3-4=-x-y+2y^2-1=0$$

初始点为(0，0)．（不要借助数值方法来执行所有计算．）

2. 考虑同伦 $h(t, x)=tf(x)+(1-t)g(x)$，其中

$$f(x)=x^2-5x+6 \qquad g(x)=x^2-1$$

证明不存在连接 g 的根与 f 的根的路径．

3. 设 $y=y(s)$ 是从 $\mathbb{R}$ 到 $\mathbb{R}^n$ 的可微函数，并且满足微分方程(9)．假定 $h(y(0))=0$．证明 $h(y(s))=0$．

4. 如果例 2 中的同伦方法用于方程组

$$\sin x+\cos y+e^{xy}=\tan^{-1}(x+y)-xy=0$$

初始点为(0，0)，试问怎样的微分方程组将决定此路径．求解的计算机程序或许会有启发．

5. 证明同伦是在从一个拓扑空间到另一个拓扑空间的连续映射之间的一种等价关系．

6. 函数 $f(x)=\sin x$ 和 $g(x)=\cos x$ 是否同伦？

7. 考虑从[0，1]到[0，1]∪[2，3]内的映射：

$$f(x)=0 \qquad g(x)=2$$

它们是否同伦？

138

第4章 解线性方程组

4.0 概述

本章中我们将构造求解问题 $Ax=b$ 的通用算法. 然后，我们分析与计算机解相关的误差并研究控制和减少误差的方法. 最后，我们对此问题介绍重要的迭代算法.

本章的总体目标是讨论求解线性方程组

$$\begin{cases} a_{11}x_1+a_{12}x_2+a_{13}x_3+\cdots+a_{1n}x_n=b_1 \\ a_{21}x_1+a_{22}x_2+a_{23}x_3+\cdots+a_{2n}x_n=b_2 \\ a_{31}x_1+a_{32}x_2+a_{33}x_3+\cdots+a_{3n}x_n=b_3 \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\vdots \\ a_{n1}x_1+a_{n2}x_2+a_{n3}x_3+\cdots+a_{nn}x_n=b_n \end{cases}$$

的数值计算方面的问题. 这是一个 n 个未知数 x_1，x_2，…，x_n 的 n 个方程的方程组. 元素 a_{ij} 和 b_i 假定为实数.

矩阵是表示方程组的有用的工具. 于是上面的线性方程组可写成

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}$$

然后我们可分别用 A，x 和 b 来表示这三个矩阵，从而方程就变成

$$Ax=b$$

4.1 矩阵代数

本节是对矩阵论基本概念的复习. 关于这个主题的进一步材料将在后续各节中需要时引进和讨论. 因为这是复习材料，大多数读者可浏览一下或跳过.

矩阵是具有如下形式的一个矩形数组：

$$\begin{bmatrix} 3.0 & 1.1 & -0.12 \\ 6.2 & 0.0 & 0.15 \\ 0.6 & -4.0 & 1.3 \\ 9.3 & 2.1 & 8.2 \end{bmatrix} \quad \begin{bmatrix} 3 & 6 & \frac{11}{7} & -17 \end{bmatrix} \quad \begin{bmatrix} 3.2 \\ -4.7 \\ 0.11 \end{bmatrix}$$

这些分别是 4×3，1×4 和 3×1 矩阵. 在描述一个矩阵的维数中，我们首先给出**行数**(水平线)，其次给出**列数**(垂直线). 一个 $m\times1$ 矩阵称为**列向量**或者就称为**向量**.

如果 A 是一个矩阵，则用记号 a_{ij}，$(A)_{ij}$ 或 $A(i,j)$ 表示第 i 行和第 j 列相交处的元素. 例如，如果 A 表示上面展示的第一个矩阵，那么 $A(3,2)=-0.4$. 矩阵 A 的**转置**用 A^{T} 表示，

它是用$(A^{T})_{ij}=a_{ji}$定义的矩阵. 利用同样的例子说明**转置**，我们有

$$A^{T}=\begin{bmatrix} 3.0 & 6.2 & 0.6 & 9.3 \\ 1.1 & 0.0 & -4.0 & 2.1 \\ -0.12 & 0.15 & 1.3 & 8.2 \end{bmatrix}$$

若矩阵 A 有性质 $A^{T}=A$，则我们就说 A 是**对称的**.

若 A 是一个矩阵而 λ 是一个纯量(在此处它是一个实数)，则用$(\lambda A)_{ij}=\lambda a_{ij}$来定义 λA. 若 $A=(a_{ij})$和 $B=(b_{ij})$是 $m\times n$ 矩阵，则用$(A+B)_{ij}=a_{ij}+b_{ij}$来定义 $A+B$. 当然，$-A$ 意味$(-1)A$. 若 A 是 $m\times p$ 矩阵，而 B 是 $p\times n$ 矩阵，则 AB 是 $m\times n$ 矩阵，用下式定义：

$$(AB)_{ij}=\sum_{k=1}^{p}a_{ik}b_{kj} \qquad (1\leqslant i\leqslant m,1\leqslant j\leqslant n)$$

下面是一些代数运算的例子：

[140]

$$\begin{bmatrix} 1 & 3 \\ 2 & -1 \\ 4 & -4 \end{bmatrix}+\begin{bmatrix} 6 & 0 \\ 3 & -7 \\ 8 & 2 \end{bmatrix}=\begin{bmatrix} 7 & 3 \\ 5 & -8 \\ 12 & -2 \end{bmatrix}$$

$$3\begin{bmatrix} 1 & 3 \\ 2 & -1 \\ 4 & -4 \end{bmatrix}=\begin{bmatrix} 3 & 9 \\ 6 & -3 \\ 12 & -12 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 1 & 3 \\ 1 & 5 & -6 \\ 2 & 1 & 5 \\ 0 & 1 & -2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ -5 & 4 \\ 0 & 3 \end{bmatrix}=\begin{bmatrix} -3 & 13 \\ -24 & 2 \\ -3 & 19 \\ -5 & -2 \end{bmatrix}$$

当我们处理线性方程组时，等价性概念是重要的. 设给出两个方程组，每一个都由 n 个未知数的 n 个方程组成：

$$Ax=b \quad Bx=d$$

若两个方程组正好有相同的解，则我们称它们为**等价的方程组**. 因此，为求解一个方程组，我们可替换求解任何的等价方程组；不丢失解并且不出现新的解. 这个简单的思想就是我们数值方法的核心. 给定一个有待求解的方程组，我们先用一些确定的初等运算把它变换成为一个更简单的等价方程组，然后替换求解.

在前段中提出的**初等运算**有下面三种类型.（这里$\mathcal{E}_i$表示方程组中第 i 个方程.）

1. 交换方程组中的两个方程：$\mathcal{E}_i\leftrightarrow\mathcal{E}_j$.
2. 用一个非零数乘一个方程：$\lambda\mathcal{E}_i\rightarrow\mathcal{E}_i$.
3. 一个方程加上某个其他方程的倍数：$\mathcal{E}_i+\lambda\mathcal{E}_j\rightarrow\mathcal{E}_i$.

·**定理 1**(等价方程组定理) 若一个方程组是由另一个方程组通过有限个初等运算得到的，则这两个方程组是等价的.

证明 考察单独应用每个初等运算的影响就足够了. 假如一个初等运算变换方程组 $Ax=b$ 为方程组 $Bx=d$. 若运算为类型 1，则这两个方程组正好由同样的方程所组成，虽然写成了不同的次序. 显然，如果 x 是第一个方程组的解，那么它也是第二个方程组的解，反之亦然. 若

运算为类型 2，则假定第 i 个方程被乘以一个纯量 λ，$\lambda \neq 0$. 在 $Ax=b$ 中第 i 个方程和第 j 个方程是

$$a_{i1}x_1 + \cdots + a_{in}x_n = b_i \tag{1}$$

和

$$a_{j1}x_1 + \cdots + a_{jn}x_n = b_j \tag{2}$$

而在 $Bx=d$ 中第 i 个方程是

$$\lambda a_{i1}x_1 + \cdots + \lambda a_{in}x_n = \lambda b_i \tag{3}$$

[141]

因为 $\lambda \neq 0$，所以任何满足方程(1)的向量 x 也满足方程(3)，反之亦然. 最后，假如运算为类型 3. 设 λ 乘第 j 个方程被加到第 i 个方程上. 则 $Bx=d$ 中的第 i 个方程为

$$(a_{i1} + \lambda a_{j1})x_1 + \cdots + (a_{in} + \lambda a_{jn})x_n = b_i + \lambda b_j \tag{4}$$

特别地观察到 $Bx=d$ 中第 j 个方程没有改变. 若 $Ax=b$，则方程(1)和(2)是正确的. 因此，方程(4)是正确的. 于是，有 $Bx=d$. 另一方面，若我们假定 x 是 $Bx=d$ 的解，则方程(4)和(2)是正确的. 若由方程(4)减去 λ 乘以方程(2)，所得结果是方程(1). 因此，$Ax=b$. ■

4.1.1 矩阵性质

$n\times n$ 矩阵

$$I = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

称为**单位矩阵**. 对任何 $n\times n$ 矩阵 A，它具有性质：$IA=A=AI$.

若 A 和 B 是两个满足 $AB=I$ 的矩阵，则我们称 B 是 A 的**右逆**，而 A 是 B 的**左逆**. 例如，

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \alpha & \beta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{5}$$

我们从此例可看出，如果一个矩阵有一个右逆，则右逆不一定是唯一的. 而对方阵来说，情况较好些，正如我们下面所指出的那样.

定理 2(右逆的定理) *一个方阵最多可能具有一个右逆.*

证明 设 $AB=I$，这里 A，B，I 都是 $n\times n$ 矩阵. 用 $A^{(j)}$ 表示 A 的第 j 列且用 $I^{(k)}$ 表示 I 的第 k 列. 等式 $AB=I$ 意味着

$$\sum_{j=1}^{n} b_{jk}A^{(j)} = I^{(k)} \qquad (1 \leqslant k \leqslant n) \tag{6}$$

所以 I 的每列都是 A 的列的线性组合. 因为 I 的列生成$\mathbb{R}^n$，所以 A 的列也都是 I 的列的线性组合. 因此，A 的列构成$\mathbb{R}^n$的一个基底，从而(6)式中的系数 b_{jk} 是唯一确定的. ■

定理 3(矩阵逆的定理) *若 A 和 B 是使得 $AB=I$ 的方阵，则 $BA=I$.*

[142]

证明 设 $C=BA-I+B$，则

$$AC = ABA - AI + AB = A - A + I = I$$

于是，C 是 A 的一个右逆（如同 B 一样）. 由定理 2，$B=C$；因此，$BA=I$. ■

由上述两个定理可得，若方阵 A 有一个右逆 B，则 B 是唯一的且 $BA=AB=I$. 于是我们称 B 为 A 的**逆**并且说 A 是**可逆的**或**非奇异的**.（当然 B 也是可逆的且 A 是它的逆.）我们记 $B=A^{-1}$ 和 $A=B^{-1}$. 下面是一个例子：

$$\begin{bmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}=\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{bmatrix}=\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

若 A 可逆，则方程组 $Ax=b$ 有解 $x=A^{-1}b$. 如果 A^{-1} 已经得到，那么这个等式提供了一个计算 x 的好方法. 如果还没有得到 A^{-1}，那么一般来说，不应该仅仅为了得到 x 计算 A^{-1}. 更高效的方法将在后面几节中讨论.

正如我们现在指出的那样，前面讨论的初等运算可以用矩阵乘法来执行. 我们把初等运算应用于 $n\times n$ 单位矩阵时产生的一个 $n\times n$ 矩阵定义为**初等矩阵**. 根据矩阵 A 的行表达的初等运算是：

1. 交换 A 中的两行：$A_s \leftrightarrow A_t$.
2. 用一个非零常数乘一行：$\lambda A_s \to A_s$.
3. 一行加另一行的倍数：$A_s + \lambda A_t \to A_s$.

在此，我们用下标 A_s，A_t 等等来表示 A 的行. 对 A 的每个初等行运算可以用一个初等矩阵左乘 A 来实现. 下面是三个例子，它们说明三类运算：

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}=\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{31} & a_{32} & a_{33} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}=\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ \lambda a_{21} & \lambda a_{22} & \lambda a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda & 1 \end{bmatrix}\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}=\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ \lambda a_{21}+a_{31} & \lambda a_{22}+a_{32} & \lambda a_{23}+a_{33} \end{bmatrix}$$

如果我们希望对 A 连续应用初等行运算，那么引入初等矩阵 E_1，E_2，…，E_m. 然后记变换后的矩阵为

143

$$E_m E_{m-1} \cdots E_2 E_1 A$$

若一个矩阵是可逆的，如此一系列初等行运算可应用于 A，化 A 为 I. 于是，我们有

$$E_m E_{m-1} \cdots E_2 E_1 A = I$$

由此可得 $A^{-1}=E_m E_{m-1}\cdots E_2 E_1$. 因此，可以通过对 I 作相同的一系列初等行运算得到 A^{-1}. 下面是说明计算逆的一个例子：

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 3 & 3 \\ 2 & 4 & 7 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = I$$

$$E_1 A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 0 \\ 2 & 4 & 7 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = E_1 I$$

$$E_2 E_1 A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} = E_2 E_1 I$$

$$E_3 E_2 E_1 A = \begin{bmatrix} 1 & 0 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 3 & -2 & 0 \\ -1 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} = E_3 E_2 E_1 I$$

$$E_4 E_3 E_2 E_1 A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 9 & -2 & -3 \\ -1 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} = E_4 E_3 E_2 E_1 I = A^{-1}$$

其中初等矩阵是

$$E_1 = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad E_3 = \begin{bmatrix} 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$E_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} \qquad E_4 = \begin{bmatrix} 1 & 0 & -3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

定理 4(非奇异矩阵性质定理) 对 $n \times n$ 矩阵 A，下列性质等价：

1. A 的逆存在；即 A 是非奇异的.
2. A 的行列式非零.
3. A 的行构成$\mathbb{R}^n$的一个基底.
4. A 的列构成$\mathbb{R}^n$的一个基底.
5. 作为$\mathbb{R}^n$到$\mathbb{R}^n$的一个映射，A 是单射的(一对一的).
6. 作为$\mathbb{R}^n$到$\mathbb{R}^n$的一个映射，A 是满射的(映上的).
7. 方程 $Ax=0$ 推出 $x=0$.
8. 对每个 $b \in \mathbb{R}^n$，刚好存在一个 $x \in \mathbb{R}^n$使得 $Ax=b$.
9. A 是初等矩阵的乘积.
10. 0 不是 A 的特征值.

144

一个重要的基本概念是矩阵的**正定性**. 如果对每个非零向量 x 有 $x^{\mathrm{T}}Ax>0$，那么称矩阵 A 是**正定的**. 例如，矩阵

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

是正定的，因为除了 $x_1=x_2=0$ 之外，对一切 x_1 和 x_2

$$x^{T}Ax=[x_1\quad x_2]\begin{bmatrix}2&1\\1&2\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}=(x_1+x_2)^2+x_1^2+x_2^2>0$$

这里 $x^{T}Ax$ 称为**二次型**. 由习题4.1.17～4.1.19可知，当处理正定性时，我们可假定对称性. 利用定义来证明一个矩阵的正定性通常不是件容易的事，因为它涉及一个任意的 $x\neq0$. 若 A 是正定和对称的，则它的特征值都是正实数.

4.1.2 分块矩阵

把矩阵分块成子矩阵并且把子矩阵看成数来计算矩阵的积通常是很方便的. 下面是这种方法的一个例子.

$$\begin{bmatrix}\begin{bmatrix}1&2\end{bmatrix}&\begin{bmatrix}1&-1&0&1\end{bmatrix}\\ \begin{bmatrix}-1&1\\0&1\\1&-1\\1&0\end{bmatrix}&\begin{bmatrix}1&0&-1&1\\-1&1&0&1\\0&0&1&0\\1&2&1&0\end{bmatrix}\end{bmatrix}\begin{bmatrix}\begin{bmatrix}1&0&1\\-1&1&2\end{bmatrix}&\begin{bmatrix}2&1\\0&1\end{bmatrix}\\ \begin{bmatrix}1&0&1\\-1&1&0\\2&1&0\\0&1&1\end{bmatrix}&\begin{bmatrix}1&2\\0&1\\-2&1\\-1&1\end{bmatrix}\end{bmatrix}$$

$$=\begin{bmatrix}\begin{bmatrix}1&2&7\end{bmatrix}&\begin{bmatrix}2&5\end{bmatrix}\\ \begin{bmatrix}-3&1&3\\-3&3&2\\4&0&-1\\2&3&2\end{bmatrix}&\begin{bmatrix}0&2\\-2&1\\0&1\\1&6\end{bmatrix}\end{bmatrix}$$

如果这些矩阵被分块成式中所指出的那样，并且用单个字母来表示子矩阵的话，那么就有下列形式的乘积

145

$$\begin{bmatrix}A_{11}&A_{12}\\A_{21}&A_{22}\end{bmatrix}\begin{bmatrix}B_{11}&B_{12}\\B_{21}&B_{22}\end{bmatrix}=\begin{bmatrix}C_{11}&C_{12}\\C_{21}&C_{22}\end{bmatrix}$$

我们可以验证 $C_{ij}=\sum_{s=1}^{2}A_{is}B_{sj}$. 例如，我们有

$$[1\quad 2]\begin{bmatrix}1&0&1\\-1&1&2\end{bmatrix}+[1\quad -1\quad 0\quad 1]\begin{bmatrix}1&0&1\\-1&1&0\\2&1&0\\0&1&1\end{bmatrix}=[1\quad 2\quad 7]$$

并且如计算所示那样，$C_{11}=A_{11}B_{11}+A_{12}B_{21}$.

为建立这种方法的一个一般的结果，设 A，B，C 是被分块成下列子矩阵的矩阵：

$$A=\begin{bmatrix}A_{11}&A_{12}&\cdots&A_{1n}\\A_{21}&A_{22}&\cdots&A_{2n}\\\vdots&\vdots&\ddots&\vdots\\A_{m1}&A_{m2}&\cdots&A_{mn}\end{bmatrix}\qquad B=\begin{bmatrix}B_{11}&B_{12}&\cdots&B_{1k}\\B_{21}&B_{22}&\cdots&B_{2k}\\\vdots&\vdots&\ddots&\vdots\\B_{n1}&B_{n2}&\cdots&B_{nk}\end{bmatrix}\qquad C=\begin{bmatrix}C_{11}&C_{12}&\cdots&C_{1k}\\C_{21}&C_{22}&\cdots&C_{2k}\\\vdots&\vdots&\ddots&\vdots\\C_{m1}&C_{m2}&\cdots&C_{mk}\end{bmatrix}$$

定理 5(分块矩阵乘法定理) 若可以构成每个乘积 $A_{is}B_{sj}$，并且若 $C_{ij}=\sum_{s=1}^{n}A_{is}B_{sj}$，则 $C=AB$.

证明 设 A_{ij} 的维数是 $m_i\times n_j$，B_{ij} 的维数是 $\hat{m}_i\times\hat{n}_j$. 因为 $A_{is}B_{sj}$ 存在，所以对一切 s，我们必须有 $n_s=\hat{m}_s$. 于是 C_{ij} 将有维数 $m_i\times\hat{n}_j$. 现在选择矩阵 C 中的任意一个元素 c_{ij}. 假如 c_{ij} 在块 C_{rs} 中，并且位于 C_{rs} 的第 p 行和第 q 列. 则我们一定有

$$i=m_1+m_2+\cdots+m_{r-1}+p \tag{7}$$

$$j=\hat{n}_1+\hat{n}_2+\cdots+\hat{n}_{s-1}+q \tag{8}$$

因此，我们有

$$c_{ij}=(C_{rs})_{pq}=\Big(\sum_{t=1}^{n}A_{rt}B_{ts}\Big)_{pq}=\sum_{t=1}^{n}(A_{rt}B_{ts})_{pq}=\sum_{t=1}^{n}\sum_{\alpha=1}^{n_t}(A_{rt})_{p\alpha}(B_{ts})_{\alpha q}$$

因为(7)式，所以元素 $(A_{rt})_{p\alpha}$ 位于 A 的第 i 行. 因为 $1\leqslant t\leqslant n$ 且 $1\leqslant\alpha\leqslant n_t$，所以这些元素填满 A 的整个 i 行. 由于(8)式，按类似的推理表明元素 $(B_{ts})_{\alpha q}$ 位于 B 的第 j 列，而且，呈现 B 的整个 j 列并以其自然次序出现. 因此，

146

$$c_{ij}=\sum_{\beta=1}^{n}(A)_{i\beta}(B)_{\beta j}=(AB)_{ij}$$ ■

习题 4.1

1. 说明第一种类型的初等运算可用 4 个其他类型的初等运算来实现，并且我们可限制 λ 为 ±1.
2. 对方程个数与未知数个数不同的方程组，定理 1 是否成立?
3. 证明：每个初等运算可以用相同类型的一个运算还原.
4. 考察线性方程组 $Ax=b$，这里 A 是 $m\times n$ 矩阵，x 是 $n\times1$ 而 b 是 $m\times1$. 用 A_1，A_2，…，A_n 表示 A 的列. 证明：方程组有解当且仅当 b 是由 $\{A_1, A_2, \cdots, A_n\}$ 线性生成的. 证明：若 $\{A_1, A_2, \cdots, A_n\}$ 线性无关，则方程组至多有一个解.
5. 设 $E(p, q, \lambda)$ 是用 λ 乘 $n\times n$ 单位阵的第 q 行加到第 p 行后得到的矩阵.（假定 $p\neq q$.）证明关系 $E(p, q, \lambda)^{-1}=E(p, q, -\lambda)$ 成立. 证明对任意的 $m\times n$ 矩阵 A，可以在 A 中把 λ 乘以第 p 列加到第 q 列算出乘积 $AE(p, q, \lambda)$.
6. **单项矩阵**是一个方阵，其中每一行和每一列正好含有一个非零元. 证明单项矩阵是非奇异的.
7. 设 A 有分块形式

$$A=\begin{bmatrix}B & C\\ 0 & I\end{bmatrix}$$

其中块是 $n\times n$ 的. 证明：若 $B-I$ 非奇异，则对 $k\geqslant1$，

$$A^k=\begin{bmatrix}B^k & (B^k-I)(B-I)^{-1}C\\ 0 & I\end{bmatrix}$$

8. (续)参照上题，当

$$A=\begin{bmatrix}B & 0\\ C & I\end{bmatrix}$$

时，求 A^k 的块结构. 并用数学归纳法证明你的结论.

9. 执行下列两个矩阵的乘法——先用块积方法，再用通常的乘法.

$$\begin{bmatrix} \begin{bmatrix} 1 & 2 \end{bmatrix} & \begin{bmatrix} 1 & -1 & 0 & 1 \end{bmatrix} \\ \begin{bmatrix} -1 & 1 \\ 0 & 1 \\ 1 & -1 \\ 1 & 0 \end{bmatrix} & \begin{bmatrix} 1 & 0 & -1 & 1 \\ -1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 2 & 1 & 0 \end{bmatrix} \end{bmatrix} \begin{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 2 \end{bmatrix} & \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix} \\ \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \\ 2 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} & \begin{bmatrix} 1 & 2 \\ 0 & 1 \\ -2 & 1 \\ -1 & 1 \end{bmatrix} \end{bmatrix}$$

10. 证明上三角 $n\times n$ 矩阵集合是全体 $n\times n$ 矩阵代数的子代数. 换言之，证明这个集合在加法、乘法和数乘运算之下是代数封闭的.

147

11. 证明：非奇异上三角矩阵之逆也是上三角矩阵.
12. 设 A 是 $n\times n$ 可逆矩阵，且设 u 和 v 是$\mathbb{R}^n$中的两个向量. 求 u 和 v 使

$$\begin{bmatrix} A & u \\ v^T & 0 \end{bmatrix}$$

可逆的必要和充分条件. 当逆存在时，给出逆矩阵的公式.

13. 设 D 是一个分块形式的矩阵：

$$D=\begin{bmatrix} A & B \\ C & I \end{bmatrix}$$

证明：若 $A-BC$ 非奇异，则 D 也非奇异.

14. (续)证明 D 的零空间维数不大于 $A-BC$ 的零空间维数这个更强的结果.
15. 下列这些矩阵是否正定?

a. $\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$

b. $\begin{bmatrix} 4 & 2 & 1 \\ 2 & 5 & 2 \\ 1 & 2 & 4 \end{bmatrix}$

16. 对 a 的什么值，下列矩阵正定?

$$A=\begin{bmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{bmatrix}$$

17. 如果 $A^T=-A$，方阵 A 称为反对称的. 证明：若 A 是反对称的，则对一切 x，$x^TAx=0$.
18. (续)证明一个反对称矩阵的对角元为 0. 并当这个矩阵是奇数阶时，证明它的行列式为 0.
19. (续)设 A 是任意方阵，定义 $A_0=(A+A^T)/2$，$A_1=(A-A^T)/2$. 证明 A_0 对称，A_1 反对称，$A=A_0+A_1$，并且对一切 x，$x^TAx=x^TA_0x$. 这说明了为什么在讨论二次型时，我们把注意力限制在对称矩阵范围内.
20. 给出一个所有元素是正的对称矩阵使得 x^TAx 有时是负的例子.
21. 一个矩阵是否可以有一个右逆和一个左逆而它们不相等?

计算机习题 4.1

1. (程序设计课题)对有兴趣从事数值实验的读者，我们提出一个程序设计课题. 它可分几个阶段实现. 这个课题涉及编写若干子程序去做线性代数中的基本作业. 这组子程序将提供求解线性方程组，用不同方法分解矩阵，计算特征值和广义逆的一个私人软件包. 相继的部分为计算机习题 4.2.1，4.2.3，4.3.1～4.3.8 和 4.4.1.
2. 对下列问题编写和测试子程序或过程：

148

a. Store(n, x, y), 用 n 维向量 x 替代 n 维向量 y: $y \leftarrow x$.

b. Prod(m, n, A, x, y), 用 $m\times n$ 矩阵 A 左乘 n 维向量 x 并把结果存放在 m 维向量 y 中: $y \leftarrow Ax$.

c. Mult(k, m, n, A, B, C), 计算 $C \leftarrow AB$, 这里 A 是 $k\times m$, B 是 $m\times n$, 而 C 是 $k\times n$ 矩阵.

d. Dot(n, x, y, a), 计算(用双精度运算)内积 $a \leftarrow \sum_{i=1}^{n} x_i y_i$, 并且存放答案为一个单精度实数 a. 注意: x_i, y_i, a 都是单精度数.

4.2 *LU* 分解和楚列斯基分解

让我们考虑一个 n 个未知数 x_1, x_2, …, x_n 的 n 个线性方程的方程组. 它可以写成如下形式

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}$$

将这个等式中的矩阵用 A, x 和 b 表示. 那么, 我们的方程组就是

$$Ax = b \tag{1}$$

4.2.1 容易求解的方程组

我们开始寻找能容易求解的特殊类型的方程组. 例如, 假如 $n\times n$ 矩阵有**对角线结构**. 这意味着 A 的所有非零元位于主对角线上, 方程组(1)是

$$\begin{bmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & \cdots & 0 \\ 0 & 0 & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}$$

此时, 我们的方程组化为 n 个简单的方程, 并且这个解是

$$x = \begin{bmatrix} b_1/a_{11} \\ b_2/a_{22} \\ b_3/a_{33} \\ \vdots \\ b_n/a_{nn} \end{bmatrix}$$

若对某个下标 i, $a_{ii}=0$ 且 $b_i=0$, 则 x_i 可以是任意实数. 若 $a_{ii}=0$ 而 $b_i\neq 0$, 则方程组无解.

149

继续搜寻容易求解的方程组(1), 我们假定 A 有**下三角结构**. 这意味着 A 的所有非零元位于主对角线上或其下方, 方程组(1)为

$$\begin{bmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ a_{21} & a_{22} & 0 & \cdots & 0 \\ a_{31} & a_{32} & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}$$

为了解这个方程组，假定对一切 i，$a_{ii} \neq 0$；然后，从第一个方程得到 x_1. 把已知的这个 x_1 值代入第二个方程，得出第二个方程的解 x_2. 用同样的方法依这个次序一次得到一个 x_i，最后可得到 x_1，x_2，…，x_n. 此时求解的一个有效的算法称为**向前回代**：

input n，(a_{ij})，(b_i)
for $i=1$ **to** n **do**

$\quad x_i \leftarrow \left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j\right) \Big/ a_{ii}$

end do
output (x_i)

习惯上，在 $\beta < \alpha$ 的情况下，形如 $\sum_{i=\alpha}^{\beta} x_i$ 的任何和式被认为是 0.

同样的思想可用来求解具有**上三角结构**的方程组. 这样的矩阵方程组具有形式

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & a_{22} & a_{23} & \cdots & a_{2n} \\ 0 & 0 & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}$$

必须再次假定：对 $1 \leqslant i \leqslant n$，$a_{ii} \neq 0$. 求解 x 的一个有效的算法如下，并且称之为**向后回代**.

input n，(a_{ij})，(b_i)
for $i=n$ **to** 1 **step** -1 **do**

$\quad x_i \leftarrow \left(b_i - \sum_{j=i+1}^{n} a_{ij}x_j\right) \Big/ a_{ii}$

end do
output (x_i)

还存在另一种利用上述思想容易求解的简单的方程组类型——即，置换一个三角方程组中

150

的方程所得到的方程组. 为了说明这点，考察方程组

$$\begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & a_{23} \\ a_{31} & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \tag{2}$$

只需对这些方程重新排序，我们就可得到一个下三角方程组：

$$\begin{bmatrix} a_{31} & 0 & 0 \\ a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_3 \\ b_1 \\ b_2 \end{bmatrix} \tag{3}$$

这个方程组就可按前面同样的方法求解．换一种方式，我们不按照通常次序 1，2，3 而按照次序 3，1，2 来求解方程组(2)中的方程．

让我们试着用正规的方式来描述前面矩阵的性质．我们假定 A 的某一行，譬如第 p_1 行，在 2，3，…，n 处为零．然后，另一行，譬如第 p_2 行，在 3，4，…，n 处为零，如此等等．如果是这种情况，我们将用第 p_1 行得到 x_1，第 p_2 行得到 x_2，…，第 p_n 行得到 x_n．若我们按次序 p_1，p_2，…，p_n 重新排列这些行，则所得的矩阵应该是下三角阵．

如果 A 是刚才考虑过的置换的上三角或下三角矩阵，如何求解 $Ax=b$ 呢？让我们假定已知或事先以某种方式确定了**置换向量**(p_1，p_2，…，p_n)．修改前面的算法．我们得出**置换的下三角方程组的向前回代**．

```
input n, (a_ij), (b_i), (p_i)
for i=1 to n do
    x_i ← (b_{p_i} − Σ_{j=1}^{i−1} a_{p_i j} x_j) / a_{p_i i}
end do
output (x_i)
```

当然，仅当 A 有性质：$a_{p_i j}=0$ $(j>i)$和 $a_{p_i i}\neq 0$(对一切 i)时，这个程序才能运行．类似地，**置换的上三角方程组的向后回代**如下：

```
input n, (a_ij), (b_i), (p_i)
for i= n to 1 step−1 do
    x_i ← (b_{p_i} − Σ_{j=i+1}^{n} a_{p_i j} x_j) / a_{p_i i}
end do
output (x_i)
```

在 $j<i$，$a_{p_i j}\neq 0$ 且对一切 i，$a_{p_i i}\neq 0$ 的情况下，这个算法起作用．

上述四个算法都没有给出直接转换成大多数程序设计语言所需的足够细节．例如，**置换的上三角方程组的向后回代算法**可能就需要更为详尽的指令： 151

```
input n, (a_ij), (b_i), (p_i)
for i=n to 1 step −1 do
    s←b_{p_i}
    for j=i+1 to n do
        s←s − a_{p_i j} x_j
    end do
    x_i←s/a_{p_i i}
end do
output (x_i)
```

4.2.2 LU 分解

假如 A 可以分解为一个下三角阵 L 和一个上三角阵 U 之积：$A=LU$．则求解方程组 $Ax=b$问题就会分成两步：

$$Lz=b \qquad \text{解 } z$$

$$Ux = z \qquad 解\ x$$

前面的分析表明求解这两个三角方程组是简单的.

我们将说明倘若在计算的某些步中不会遇到 0 除数，怎样实现分解 $A=LU$. 并不是每个矩阵都有这样的分解，现在将研究这个困难.

我们用一个 $n\times n$ 矩阵 A 开始搜索矩阵

$$L = \begin{bmatrix} \ell_{11} & 0 & 0 & \cdots & 0 \\ \ell_{21} & \ell_{22} & 0 & \cdots & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \cdots & \ell_{nn} \end{bmatrix}$$

$$U = \begin{bmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & u_{2n} \\ 0 & 0 & u_{33} & \cdots & u_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & u_{nn} \end{bmatrix}$$

使得

$$A = LU \tag{4}$$

如果这是可能的话，我们就说 A 有一个 **LU 分解**. 由(4)式可证明 L 和 U 不是唯一确定的. 事实上，对每个 i，我们可以对 ℓ_{ii} 或 u_{ii}(但不是两者)指定一个非零值. 例如，一个简单的选择是取 $\ell_{ii}=1$，$i=1, 2, \cdots, n$. 于是使 L 为**单位下三角阵**. 另一个明显的选择使 U 为**单位上三角阵**(对每个 i，$u_{ii}=1$). 这些特殊情况具有特别的重要性.

152

为导出 A 的 LU 分解的算法，我们从矩阵乘法公式出发：

$$a_{ij} = \sum_{s=1}^{n} \ell_{is} u_{sj} = \sum_{s=1}^{\min(i,j)} \ell_{is} u_{sj} \tag{5}$$

这里我们利用了事实：$s>i$ 时 $\ell_{is}=0$ 且 $s>j$ 时 $u_{sj}=0$.

这个过程的每一步确定 U 的一个新行和 L 中一个新列. 在第 k 步，我们可以假定 U 的第 $1, 2, \cdots, k-1$ 行和 L 的第 $1, 2, \cdots, k-1$ 列已算出. (若 $k=1$，这个假设无意义.)在(5)式中取 $i=j=k$，我们得到

$$a_{kk} = \sum_{s=1}^{k-1} \ell_{ks} u_{sk} + \ell_{kk} u_{kk} \tag{6}$$

若 u_{kk} 或 ℓ_{kk} 已指定，我们利用(6)式来确定另一个. 已知 ℓ_{kk} 和 u_{kk} 时，我们用(5)式分别写出第 k 行($i=k$)和第 k 列($j=k$)，

$$a_{kj} = \sum_{s=1}^{k-1} \ell_{ks} u_{sj} + \ell_{kk} u_{kj} \qquad (k+1 \leqslant j \leqslant n) \tag{7}$$

$$a_{ik} = \sum_{s=1}^{k-1} \ell_{is} u_{sk} + \ell_{ik} u_{kk} \qquad (k+1 \leqslant i \leqslant n) \tag{8}$$

若 $\ell_{kk}\neq 0$，就可以用(7)式来得到元素 u_{kj}. 类似地，若 $u_{kk}\neq 0$，就可用(8)式来得到元素 ℓ_{ik}.

有趣的是这两个计算理论上可以并行实现(即同时地). 在某些计算机上执行时间内用相当大的存储量这件事情实际上是可行的. 细节见 Kincaid and Oppe[1988]. 正如所述的那样计算 U 的第 k 行和 L 的第 k 列完成算法的第 k 步. 而这些计算需要用 ℓ_{kk} 和 u_{kk} 作除法；因此，若这些除数为 0，则计算通常不能完成. 然而，在某些情况下，它们又是可以完成的. (见习题 4.2.43.)

基于前面的分析，当 L 是单位下三角阵($\ell_{ii}=1$, $1\leqslant i\leqslant n$)时，算法称为 **Doolittle 分解**. 当 U 是单位上三角阵($u_{ii}=1$, $1\leqslant i\leqslant n$)时，算法称为**克劳特分解**. 当 $U=L^{T}$ 使得 $\ell_{ii}=u_{ii}$, $1\leqslant i\leqslant n$，算法称为**楚列斯基分解**. 我们将在本节的后面详细讨论楚列斯基分解，因为这个分解要求 A 有若干特殊的性质：即 A 应该是实对称和正定的.

这些分解中哪个比较好呢？因为每个分解都与基本的高斯消元法的不同变形有关. 所以我们需要对它们有一个全面的理解.

一般的 *LU* 分解算法如下：

153

input n, (a_{ij})
for $k=1$ **to** n **do**
  Specify a nonzero value for either
    ℓ_{kk} or u_{kk} and compute the other from
    $\ell_{kk}u_{kk}=a_{kk}-\sum_{s=1}^{k-1}\ell_{ks}u_{sk}$
  for $j=k+1$ **to** n **do**
    $u_{kj}\leftarrow\left(a_{kj}-\sum_{s=1}^{k-1}\ell_{ks}u_{sj}\right)/\ell_{kk}$
  end do
  for $i=k+1$ **to** n **do**
    $\ell_{ik}\leftarrow\left(a_{ik}-\sum_{s=1}^{k-1}\ell_{is}u_{sk}\right)/u_{kk}$
  end do
end do
output (ℓ_{ij}), (u_{ij})

注意在计算 U 的第 k 行和 L 的第 k 列时可能的并行性(同时性)(见习题 4.2.56.)

详细地修改像 LU 分解和楚列斯基分解那样的算法，使得它们能以不同形式适用于高性能计算机的高精度计算. 为求得利用多指令多数据(MIMD)分布式存储并行计算机独有特性的可升级算法，在重新设计的过程中使用了很多技巧. 例如，块-分段算法被用于减少存储结构中不同层次之间数据移动的频率. 此外，特别分散版本的基本线性代数子程序(BLAS)被用作计算和通信的标准部件. 诸如 LAPACK 和 ScaLAPACK 这样的软件库也利用这些设计特性来编写. 关于这些问题的讨论可见 Anderson et al. [1995]的《LAPACK Users' Guide》与 Dongarra and Walker [1995].

在分解 $A=LU$ 中，L 是下三角阵且 U 是上三角阵. 从(5)式得到 L 和 U 的 n^2+n 个未知数的 n^2 个方程. 显然必须指定它们中的 n 个未知数. 甚至一个更一般的分解也允许指定 L 和 U 的任意 n 个元素并求解所得的方程组. 遗憾地，这个方程组可能关于 ℓ_{ij} 和 u_{ij} 是**非线性的**. (见习题 4.2.50.) 在前面的算法中，必须先指明 ℓ_{kk} 或 u_{kk}，然后在转移到下一列和下一行之前

计算 L 的第 k 列和 U 的第 k 行的全部元素，而且在这些计算中次序 $k=1$，2，…，n 是相当重要的.

154

执行 **Doolittle 分解**的伪代码如下：

```
input n, (a_ij)
for k=1 to n do
    ℓ_kk ← 1
    for j=k to n do
        u_kj ← a_kj − Σ_{s=1}^{k−1} ℓ_ks u_sj
    end do
    for i=k+1 to n do
        ℓ_ik ← (a_ik − Σ_{s=1}^{k−1} ℓ_is u_sk) / u_kk
    end do
end do
output (ℓ_ij), (u_ij)
```

例 1 求下列矩阵的 Doolittle、克劳特和楚列斯基分解

$$A=\begin{bmatrix}60&30&20\\30&20&15\\20&15&12\end{bmatrix}$$

解 从算法可得 Doolittle 分解是

$$A=\begin{bmatrix}1&0&0\\\frac{1}{2}&1&0\\\frac{1}{3}&1&1\end{bmatrix}\begin{bmatrix}60&30&20\\0&5&5\\0&0&\frac{1}{3}\end{bmatrix}\equiv LU$$

与其直接计算下面的两个分解，倒不如从上面的 Doolittle 分解得到它们. 取 U 的对角元为对角阵 D，我们可记

$$A=\begin{bmatrix}1&0&0\\\frac{1}{2}&1&0\\\frac{1}{3}&1&1\end{bmatrix}\begin{bmatrix}60&0&0\\0&5&0\\0&0&\frac{1}{3}\end{bmatrix}\begin{bmatrix}1&\frac{1}{2}&\frac{1}{3}\\0&1&1\\0&0&1\end{bmatrix}\equiv LD\hat{U}$$

通过取 $\hat{L}=LD$，我们得到克劳特分解

$$A=\begin{bmatrix}60&0&0\\30&5&0\\20&5&\frac{1}{3}\end{bmatrix}\begin{bmatrix}1&\frac{1}{2}&\frac{1}{3}\\0&1&1\\0&0&1\end{bmatrix}\equiv\hat{L}\hat{U}$$

通过分裂 $LD\hat{U}$ 分解中的 D 为形式 $D^{1/2}D^{1/2}$，并与 L 结合一个因子以及与 $\hat{U}$ 结合另一个因子得到楚列斯基分解. 于是

$$A=\begin{bmatrix}1&0&0\\ \frac{1}{2}&1&0\\ \frac{1}{3}&1&1\end{bmatrix}\begin{bmatrix}\sqrt{60}&0&0\\ 0&\sqrt{5}&0\\ 0&0&\frac{1}{3}\sqrt{3}\end{bmatrix}\begin{bmatrix}\sqrt{60}&0&0\\ 0&\sqrt{5}&0\\ 0&0&\frac{1}{3}\sqrt{3}\end{bmatrix}\begin{bmatrix}1&\frac{1}{2}&\frac{1}{3}\\ 0&1&1\\ 0&0&1\end{bmatrix}$$

$$=\begin{bmatrix}\sqrt{60}&0&0\\ \frac{1}{2}\sqrt{60}&\sqrt{5}&0\\ \frac{1}{3}\sqrt{60}&\sqrt{5}&\frac{1}{3}\sqrt{3}\end{bmatrix}\begin{bmatrix}\sqrt{60}&\frac{1}{2}\sqrt{60}&\frac{1}{3}\sqrt{60}\\ 0&\sqrt{5}&\sqrt{5}\\ 0&0&\frac{1}{3}\sqrt{3}\end{bmatrix}\equiv\tilde{L}\tilde{L}^{T}$$

■

我们最关注的分解是 $A=LU$，这里 L 是单位下三角阵而 U 是上三角阵. 因此，每当提到 LU 分解时，我们意指其中 L 是单位下三角阵. 下面是方阵 A 有 LU 分解的一个充分条件.

定理 1（LU 分解定理） 若 $n\times n$ 矩阵 A 的 n 个前主子式非奇异，则 A 有 LU 分解.

证明 回想矩阵 A 的第 k 个前主子式是矩阵

$$A_k=\begin{bmatrix}a_{11}&a_{12}&\cdots&a_{1k}\\ a_{21}&a_{22}&\cdots&a_{2k}\\ \vdots&\vdots&\ddots&\vdots\\ a_{k1}&a_{k2}&\cdots&a_{kk}\end{bmatrix}$$

设 A_k，L_k，U_k 分别是 A，L，U 的第 k 个前主子式，A_1，A_2，…，A_n 是非奇异的. 为归纳证明起见，假定已经得到 L_{k-1} 和 U_{k-1}. 在(5)式中，若 i 和 j 在范围 1，2，…，$k-1$ 中，则 s 也在这范围中. 因此，(5)式说明

$$A_{k-1}=L_{k-1}U_{k-1}$$

因为由假设 A_{k-1} 非奇异，所以 L_{k-1} 和 U_{k-1} 也非奇异. 因为 L_{k-1} 非奇异，所以，我们可解方程组

$$\sum_{s=1}^{k-1}\ell_{is}u_{sk}=a_{ik}\qquad(1\leqslant i\leqslant k-1)$$

求 u_{sk}，$1\leqslant s\leqslant k-1$. 这些元素位于 U 的第 k 列. 因为 U_{k-1} 也非奇异，所以我们可解方程组

$$\sum_{s=1}^{k-1}\ell_{ks}u_{sj}=a_{kj}\qquad(1\leqslant j\leqslant k-1)$$

求 ℓ_{ks}，$1\leqslant s\leqslant k-1$. 这些元素位于 L 的第 k 行. 因为 ℓ_{kk} 已指定为 1，所以从要求

$$a_{kk}=\sum_{s=1}^{k}\ell_{ks}u_{sk}=\sum_{s=1}^{k-1}\ell_{ks}u_{sk}+\ell_{kk}u_{kk}$$

可得到 u_{kk}. 于是，构成 L_k 和 U_k 一切必要的新元素已经确定了. 注意 $\ell_{11}u_{11}=a_{11}$，所以，$\ell_{11}=1$，$u_{11}=a_{11}$，至此归纳法完成. ■

155~156

4.2.3 楚列斯基分解

正如本节前面提到的那样，由数学家 André Louis Cholesky 名字命名的一个矩阵分解在某些情况下是有用的，楚列斯基证明了下列结果.

定理 2(楚列斯基 LL^{T} 分解定理) 若 A 是一个实对称正定阵，则它有唯一的分解 $A=LL^{\mathrm{T}}$，其中 L 是具有正对角元的下三角阵.

证明 记得如果 $A=A^{\mathrm{T}}$ 并且对每个非零向量 x 有 $x^{\mathrm{T}}Ax>0$，就称矩阵 A 为**对称正定的**. 因此即得 A 是非奇异的，因为 A 显然不能映射任何非零向量为 0 向量. 进而，考虑特别形式的向量 $x=(x_1, x_2, \cdots, x_k, 0, 0, \cdots, 0)^{\mathrm{T}}$. 可看出 A 的前主子式也是正定的. 定理 1 推出 A 有 LU 分解. 由 A 的对称性，我们有

$$LU = A = A^{\mathrm{T}} = U^{\mathrm{T}}L^{\mathrm{T}}$$

这可推得

$$U(L^{\mathrm{T}})^{-1} = L^{-1}U^{\mathrm{T}}$$

这个等式的左边是上三角阵，而右边是下三角阵.（见习题 4.2.1.）因此，存在一个对角阵 D 使 $U(L^{\mathrm{T}})^{-1}=D$. 因此，$U=DL^{\mathrm{T}}$，$A=LDL^{\mathrm{T}}$. 由习题 4.2.26 知 D 是正定的，并且它的对角元 d_{ii} 是正的. 用 $D^{1/2}$ 表示对角元是$\sqrt{d_{ii}}$的对角阵，我们有 $A=\tilde{L}\tilde{L}^{\mathrm{T}}$，其中 $\tilde{L}=LD^{1/2}$，这就是楚列斯基分解. 唯一性的证明留作习题. ■

楚列斯基分解算法是一般的 LU 分解算法的一种特殊情况. 若 A 实对称正定，则由定理 2 知，它有唯一的形如 $A=LL^{\mathrm{T}}$ 的分解，其中 L 是有正对角元的下三角阵. 因此，由(4)式，$U=L^{\mathrm{T}}$. 在一般算法的第 k 步，对角元由下式计算

$$\ell_{kk} = \left(a_{kk} - \sum_{s=1}^{k-1}\ell_{ks}^2\right)^{1/2} \tag{9}$$

157 于是，**楚列斯基分解**的算法如下：

input $n, (a_{ij})$
for $k=1$ **to** n **do**
  $\ell_{kk} \leftarrow \left(a_{kk} - \sum_{s=1}^{k-1}\ell_{ks}^2\right)^{1/2}$
  for $i=k+1$ **to** n **do**
    $\ell_{ik} \leftarrow \left(a_{ik} - \sum_{s=1}^{k-1}\ell_{is}\ell_{ks}\right)/\ell_{kk}$
  end do
end do
output (ℓ_{ij})

定理 2 保证 $\ell_{kk}>0$. 观察(9)式，对 $j\leqslant k$，给出下列界限：

$$a_{kk} = \sum_{s=1}^{k}\ell_{ks}^2 \geqslant \ell_{kj}^2$$

由此，我们得出

$$|\ell_{kj}| \leqslant \sqrt{a_{kk}} \qquad (1 \leqslant j \leqslant k)$$

因此，L 的任何元素以 A 相应对角元的平方根为界. 这就推出即使不选任何主元素，L 的元素相对于 A 也不会变大.（选主元在下节中说明.）

在楚列斯基和 Doolittle 这两个算法中，应该以双精度计算内积以避免舍入误差增大.（见计算机习题 2.2.6.）

习题 4.2

1. 证明定理 2 的证明中所需要的事实.
   a. 若 U 是可逆上三角阵，则 U^{-1} 是上三角阵.
   b. 单位下三角阵之逆是单位下三角阵.
   c. 两个上(下)三角阵之积是上(下)三角阵.
2. 证明：若非奇异阵 A 有 LU 分解，其中 L 是单位下三角阵，则 L 和 U 唯一.
3. 若 A 非奇异. 证明：向前回代和向后回代算法及它们的置换形式总能求解 $Ax=b$.
4. (续)计算这四种算法中包含的算术运算次数.
5. 证明一个上三角阵或下三角阵是非奇异的当且仅当它的对角元全不为 0.
6. 证明：若 A 的全部主子式都非奇异并且对每个 i，$\ell_{ii}\neq 0$，则对 $1\leqslant k\leqslant n$，$u_{kk}\neq 0$.

158

7. 证明 $A=\begin{bmatrix}0 & 1\\ 1 & 1\end{bmatrix}$ 没有 LU 分解. 警告：这不是本节所证明的定理 1 的一个简单推论.
8. a. 写出 **Doolittle 算法的行形式**，在第 k 步计算 L 的第 k 行和 U 的第 k 行.（因此，在第 k 步的计算次序是 ℓ_{k1}，ℓ_{k2}，…，$\ell_{k,k-1}$，u_{kk}，…，u_{kn}.）
   b. 写出 **Doolittle 算法的列形式**，在第 k 步计算 U 的第 k 列和 L 的第 k 列.（因此，在第 k 步的计算次序是 u_{1k}，u_{2k}，…，u_{kk}，$\ell_{k+1,k}$，…，$\ell_{n,k}$.）
9. 利用等式 $UU^{-1}=I$，得出一个求上三角阵之逆的算法. 假定 U^{-1} 存在；即 U 的对角元全不为 0.
10. 矩阵 $A=(a_{ij})$：当 $j>i$ 或 $j<i-1$ 时 $a_{ij}=0$，称 A 为**斯蒂尔切斯矩阵**. 设计一个求这种矩阵之逆的有效算法.
11. 设 A 是 $n\times n$ 矩阵.（p_1，p_2，…，p_n）是（1，2，…，n）的一个置换，它使得 A 中第 i 行(对 $i=1$，2，…，n)仅包含 p_1，p_2，…，p_i 列中的非零元. 编写一个求解 $Ax=b$ 的算法.
12. 说明每个形如 $A=\begin{bmatrix}0 & a\\ 0 & b\end{bmatrix}$ 的矩阵有 LU 分解. 说明即使 L 是单位下三角阵，这个分解也不唯一.（这个问题和下面两个问题，都说明了 Taussky 准则：如果关于矩阵的猜测不成立，通常可以用一个 2×2 矩阵证明其不成立.）
13. (续)说明每个形如 $A=\begin{bmatrix}0 & 0\\ a & b\end{bmatrix}$ 的矩阵有 LU 分解. 它是否有一个 LU 分解，其中的 L 是单位下三角阵?
14. (续)说明下列形式的每个矩阵有 LU 分解. 它是否有一个 LU 分解，其中的 L 是单位下三角阵?
    a. $A=\begin{bmatrix}a & 0\\ b & 0\end{bmatrix}$
    b. $A=\begin{bmatrix}a & b\\ 0 & 0\end{bmatrix}$
15. 证明：若 A 可逆且有 LU 分解，则 A 的所有主子式都非奇异.
16. 设方程组 $Ax=b$ 有下列性质：存在（1，2，…，n）的两个置换 $p=(p_1，p_2，\cdots，p_n)$ 和 $q=(q_1，q_2，\cdots，q_n)$ 使得对每个 i，编号 p_i 的方程只含有变量 x_{q_1}，x_{q_2}，…，x_{q_i}. 编写一个求解此方程组的有效算法.
17. 计算对一个单位下三角阵求逆所需要的乘法和/或除法次数.
18. 证明或否定：若 A 有一个 LU 分解，其中 L 是单位下三角阵，则 A 还有一个 LU 分解，其中 U 是单位上三角阵.
19. 假如 A 的 LU 分解已知，给出一个求 A 的逆的算法.（利用上面的习题 4.2.9 和计算机习题 4.2.1.）
20. 若矩阵 A 有性质：$a_{ij}=0$，$i+j\leqslant n$. 讨论求 A 的逆的算法.

[159] 21. 利用楚列斯基定理证明对称矩阵 A 的下列两条性质等价.

a. A 是正定的.

b. 在$\mathbb{R}^n$中存在一组线性无关的向量 $x^{(1)}$，$x^{(2)}$，…，$x^{(n)}$使得 $A_{ij}=(x^{(i)})^{T}(x^{(j)})$.

22. 证实下列求解 $Ux=b$ 算法的正确性，此时 U 是上三角阵.

```
for j=n to 1 step -1 do
    x_j ← b_j/u_jj
    for i=1 to j-1 do
        b_i ← b_i - u_ij x_j
    end do
end do
```

23. 证明：若 A 的全部前主子式非奇异，则 A 有分解 LDU，其中 L 是单位下三角阵，U 是单位上三角阵，而 D 是对角阵.

24.(续)若 A 是对称阵且其前主子式非奇异，则 A 有分解 LDL^{T}，其中 L 是单位下三角阵，而 D 是对角矩阵.

25.(续)编写一个计算对称阵 A 的 LDL^{T} 分解的算法. 你的算法应该大约做标准高斯算法一半的工作量. 注意：若 A 有奇异主子式，这个算法可能会失败.（这个修改楚列斯基算法不包含平方根计算.）

26. 证明：A 正定且 B 非奇异当且仅当 BAB^{T} 正定.

27. 若 A 正定，是否可得 A^{-1}也正定？

28. 考虑

$$A=\begin{bmatrix}2 & 6 & -4\\ 6 & 17 & -17\\ -4 & -17 & -20\end{bmatrix}$$

不用高斯消元法直接求分解 $A=LDL^{T}$，其中 D 是对角阵，L 是单位下三角阵.

29. 讨论直接求 A 的 UL 分解的算法，这里 L 是单位下三角阵，而 U 是上三角阵. 给出一个求解 $ULx=b$ 的算法.

30. 求矩阵

$$A=\begin{bmatrix}3 & 0 & 1\\ 0 & -1 & 3\\ 1 & 3 & 0\end{bmatrix}$$

的 LU 分解，其中 L 是下三角阵，而 U 是单位上三角阵.

31. 分解矩阵 $A=\begin{bmatrix}1 & 2\\ 2 & 5\end{bmatrix}$使 $A=LL^{T}$，这里 L 是下三角阵.

32. 直接求下列矩阵 A 的 LL^{T} 分解，其中 L 是具有正对角元的下三角阵.

[160]

$$A=\begin{bmatrix}4 & \frac{1}{2} & 1\\ \frac{1}{2} & \frac{17}{16} & \frac{1}{4}\\ 1 & \frac{1}{4} & \frac{33}{64}\end{bmatrix}$$

33. 假如非奇异阵 A 有楚列斯基分解. 那么 A 的行列式会怎样呢？

34. 求矩阵 $A=\begin{bmatrix}1 & 5\\ 3 & 16\end{bmatrix}$的 LU 分解，其中 L 和 U 都有单位对角元. 再把数 16 变成 15，重做一遍.

35. 考虑对称三对角正定矩阵

$$A=\begin{bmatrix}136.01 & 90.860 & 0.0 & 0.0\\ 90.860 & 98.810 & -67.590 & 0.0\\ 0.0 & -67.590 & 132.01 & 46.260\\ 0.0 & 0.0 & 46.260 & 177.17\end{bmatrix}$$

使用5个有效数字，按下列方法分解 A：

a. $A=LU$，这里 L 是单位下三角阵，U 是上三角阵.

b. $A=LDU$，这里 L 是单位下三角阵，D 是对角阵，U 是单位上三角阵.

c. $A=LU$，这里 L 是下三角阵，U 是单位上三角阵.

d. $A=LL^{\mathrm{T}}$，这里 L 是下三角阵.

36. 求矩阵

$$A=\begin{bmatrix}6 & 10 & 0\\ 12 & 26 & 4\\ 0 & 9 & 12\end{bmatrix}$$

的 LU 分解，其中 L 是主对角元为2的下三角阵.

37. 证明或否定：若一个奇异阵有 Doolittle 分解，则分解不唯一.

38. 证明分解 $A=LL^{\mathrm{T}}$ 的唯一性，其中 L 是具有正对角元的下三角阵.

39. 对称正定(SPD)的矩阵 A 有 SPD 的平方根 X. 于是 $X^2=A$. 若 $A=\begin{bmatrix}13 & 10\\ 10 & 17\end{bmatrix}$，求 X.

40. 讨论下列两种特殊情况下求解线性方程组 $Ax=b$ 的算法.

a. $a_{ij}=0$，$j\leqslant n-i$

b. $a_{ij}=0$，$j>n+1-i$

41. 利用(6)，(7)和(8)式，求矩阵

$$A=\begin{bmatrix}2 & 1 & -2\\ 4 & 2 & -1\\ 6 & 3 & 11\end{bmatrix}$$

的所有 Doolittle 分解. 在此例中，尽管 $u_{22}=0$，算法还是能执行.

42. 证明：若 A 对称，则在它的 LU 分解中 L 的列是 U 的行的倍数.

43. 对 $A=\begin{bmatrix}1 & 5\\ 3 & 17\end{bmatrix}$，求一切 LU 分解和 UL 分解，其中 L 是单位下三角阵.

161

44. 定义 **P 矩阵**，其中 $a_{ij}=0$，$j\leqslant n-i$，**Q 矩阵**是一个 P 矩阵，其中 $a_{i,n-i+1}=1$，$i=1,2,\cdots,n$. 求 $A=\begin{bmatrix}3 & 15\\ -1 & -1\end{bmatrix}$的 **$PQ$ 分解**.

45. (续)设计一个求已知矩阵 PQ 分解的算法，并且设计一个求解形如 $PQx=b$ 的方程组的算法.

46. 假定 A 的 LU 分解是可得到的，编写一个求解方程 $x^{\mathrm{T}}A=b^{\mathrm{T}}$ 的算法.

47. 如果 A 有 Doolittle 分解，那么 A 的行列式的简单公式是什么？

48. 设

$$A=\begin{bmatrix}25 & 0 & 0 & 0 & 1\\ 0 & 27 & 4 & 3 & 2\\ 0 & 54 & 58 & 0 & 0\\ 0 & 108 & 116 & 0 & 0\\ 100 & 0 & 0 & 0 & 24\end{bmatrix}$$

确定 A 的最一般的 LU 分解，其中 L 是单位下三角阵. 说明 Doolittle 算法产生一个这样的 LU 分解.

49. 设 A 是对称阵，其前主子式非奇异. 那么对 $\epsilon>0$，矩阵 $A+\epsilon I$ 具有这种相同的性质吗?
50. 考虑 2×2 矩阵 A 的 LU 分解，说明若指定 ℓ_{22} 和 u_{22}，则确定 L 和 U 的其余元素的方程是非线性的.
51. 证明：若 A 对称非负定，则对某个下三角阵 L，$A=LL^T$. 术语**非负定**意指对一切 x，$x^TAx\geqslant 0$.
52. 求 a、b、c 的确切条件使矩阵 $\begin{bmatrix} a & b \\ b & c \end{bmatrix}$ 非负定.
53. 证明：若矩阵 $\begin{bmatrix} a & b \\ b & c \end{bmatrix}$ 非负定，则它有分解 LL^T，其中 L 是下三角阵.
54. 证明或否定：对称阵非负定当且仅当它的一切前主子式有非负的行列式.
55. 求 a，b，c 使矩阵 $\begin{bmatrix} a & b \\ b & c \end{bmatrix}$ 有分解 LL^T 的充分必要条件，其中 L 是下三角阵.
56. 在本题中，使用记号 $X_{i:j,k}$ 表示 X 第 k 列的第 i 个到第 j 个元素组成的部分. 类似地，设 $X_{k,i:j}$ 表示 X 第 k 行的第 i 个到第 j 个元素组成的部分.

   a. 参考(7)式并说明它可被写成
   $$U_{k,k+1:n}=(A_{k,k+1:n}-L_{k,1:k-1}M)/\ell_{kk}$$
   其中 M 是行向量为 $U_{i,k+1:n}$，$1\leqslant i\leqslant k-1$ 的矩阵.

   b. 对(8)式执行类似的变换.

   本题所讨论的计算具有形式 $y\leftarrow y-Mx$. 它们可在向量超级计算机上非常有效地执行.（细节见 Kincaid and Oppe[1988]与 Oppe and Kincaid[1988].）

162

计算机习题 4.2

1. 设计一个求 $n\times n$ 下三角阵 A 的逆的有效算法. 建议：利用 A^{-1} 也是下三角阵这个事实. 对你的算法进行编程并且对元素为 $a_{ij}=(i+j)^2(i\geqslant j)$ 的矩阵测试这个程序. 取 $n=10$，用 AA^{-1} 的积来测试所求的逆.
2. 用楚列斯基方法解下列方程组
$$\begin{cases} 0.05x_1+0.07x_2+0.06x_3+0.05x_4=0.23 \\ 0.07x_1+0.10x_2+0.08x_3+0.07x_4=0.32 \\ 0.06x_1+0.08x_2+0.10x_3+0.09x_4=0.33 \\ 0.05x_1+0.07x_2+0.09x_3+0.10x_4=0.31 \end{cases}$$
3. 编写实现一般的 LU 分解算法的子程序或过程. 指定的对角元可存放在数组 D 中. 一个相应的逻辑数组可用于指示 D 的一个元素是否属于 L 或 U 的对角线. 用元素是 $a_{ij}=(i+j-1)^{-1}$ 的某些希尔伯特矩阵来测试这个程序. 对每个矩阵加上一个或多个其他指定的对角元来产生 Doolittle、克劳特和楚列斯基分解.

4.3 选主元和构造算法

在 4.2 节中，假借矩阵的 LU 分解给出了抽象形式上的高斯消元法. 在本节中，将描述传统形式的高斯消元法以及与抽象形式的关系. 然后，我们将对这个过程作必要的修改产生一个切实可行的计算机程序. 在这个讨论中，我们将交替使用矩阵方程组的方程和行这些词.

当高斯消元法可行时为什么我们要作 Doolittle、克劳特和楚列斯基分解呢? 在台式计算机时代，这些方法中的一个或几个可能存在超过其他方法的优势. 但随着计算机和数学软件的发展，这些微弱的优势已消失. 因此，讨论 Doolittle 和克劳特方法主要是由于历史上的原因. 另一方面，楚列斯基方法特别适合于对称正定方程组.

4.3.1 基本的高斯消元法

下面是一个用于说明高斯算法的 4 个未知数 4 个方程的简单方程组：

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 12 & -8 & 6 & 10 \\ 3 & -13 & 9 & 3 \\ -6 & 4 & 1 & -18 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}=\begin{bmatrix} 12 \\ 34 \\ 27 \\ -38 \end{bmatrix} \tag{1}$$

在过程的第 1 步，我们从第 2 个方程中减去 2 倍的第 1 个方程．然后从第 3 个方程中减去 1/2 倍的第 1 个方程．最后从第 4 个方程中减去 -1 倍的第 1 个方程．数 2，1/2，-1 称为消元过程第 1 步中的**乘子**．数 6 用于构成每个乘子的除数，称之为这一步的**主元素**．完成第 1 步计算后，方程组变成下式：

163

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & -12 & 8 & 1 \\ 0 & 2 & 3 & -14 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}=\begin{bmatrix} 12 \\ 10 \\ 21 \\ -26 \end{bmatrix} \tag{2}$$

虽然第 1 行在过程中用过，但它没有改变．在第 1 步，我们称第 1 行为**主行**．在过程的下一步，第 2 行用作主行而 -4 为主元素．我们从第 3 行中减去 3 倍的第 2 行，从第 4 行中减去 $-1/2$ 倍的第 2 行．这步的乘子是 3 和 $-\frac{1}{2}$．结果是

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 4 & -13 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}=\begin{bmatrix} 12 \\ 10 \\ -9 \\ -21 \end{bmatrix} \tag{3}$$

最后一步从第 4 行中减去 2 倍的第 3 行，故乘子和主元素都为 2．所得的方程组为

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 0 & -3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}=\begin{bmatrix} 12 \\ 10 \\ -9 \\ -3 \end{bmatrix} \tag{4}$$

这个方程组是上三角的，并且在两个方程组解相同的意义下，它**等价**于原来的方程组．求解最后的方程组很容易，只要从第 4 行出发并按行反向次序操作即可．解是

$$x=\begin{bmatrix} 1 \\ -3 \\ -2 \\ 1 \end{bmatrix}$$

用于变换方程组的乘子可以按单位下三角阵 $L=(\ell_{ij})$ 方式建立：

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ \frac{1}{2} & 3 & 1 & 0 \\ -1 & -\frac{1}{2} & 2 & 1 \end{bmatrix} \tag{5}$$

注意，每个乘子被写在矩阵中相应它负责产生 0 元素的位置上．最后的方程组的系数阵是上三

164

角阵 $U=(u_{ij})$：

$$U = \begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 0 & -3 \end{bmatrix} \tag{6}$$

这两个矩阵给出 A 的 LU 分解，这里 A 是原方程组的系数矩阵．因此，

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 12 & -8 & 6 & 10 \\ 3 & -13 & 9 & 3 \\ -6 & 4 & 1 & -18 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ \frac{1}{2} & 3 & 1 & 0 \\ -1 & -\frac{1}{2} & 2 & 1 \end{bmatrix} \begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 0 & -3 \end{bmatrix} \tag{7}$$

要看出这为什么必定成立是不困难的．若我们知道 U 是怎样从 A 得到的，则通过相反的过程可从 U 得到 A．若我们用 A_1，A_2，A_3，A_4 表示 A 的行并用 U_1，U_2，U_3，U_4 表示 U 的行，则由消元过程给出，例如，$U_2=A_2-2A_1$．因此，$A_2=2A_1+U_2=2U_1+U_2$．系数 2 和 1 占据 L 的第 2 行．类似地，行运算导致第 3 行是 $U_3=(A_3-A_1/2)-3U_2$，最后，我们有 $A_3=A_1/2+3U_2+U_3=U_1/2+3U_2+U_3$．所以系数 1/2，3，1 必须占据 L 的第 3 行，等等．

为正式地描述高斯算法过程，我们把它理解为一个连续 $n-1$ 步主步产生的如下矩阵系列：

$$A = A^{(1)} \to A^{(2)} \to \cdots \to A^{(n)}$$

在第 $k-1$ 步结束，矩阵 $A^{(k)}$ 将被构造出来；它的样子如下所示，其中用线框起来的第 k 行和仅仅前面有线的第 k 列说明由消元过程产生的结构．

$$\left[\begin{array}{ccc|ccccc}
a_{11}^{(k)} & \cdots & a_{1,k-1}^{(k)} & a_{1k}^{(k)} & \cdots & a_{1j}^{(k)} & \cdots & a_{1n}^{(k)} \\
\vdots & \ddots & & \vdots & & \vdots & & \vdots \\
0 & \cdots & a_{k-1,k-1}^{(k)} & a_{k-1,k}^{(k)} & \cdots & a_{k-1,j}^{(k)} & \cdots & a_{k-1,n}^{(k)} \\
\hline
0 & \cdots & 0 & a_{kk}^{(k)} & \cdots & a_{kj}^{(k)} & \cdots & a_{kn}^{(k)} \\
\hline
0 & \cdots & 0 & a_{k+1,k}^{(k)} & \cdots & a_{k+1,j}^{(k)} & \cdots & a_{k+1,n}^{(k)} \\
\vdots & & \vdots & \vdots & & \vdots & & \vdots \\
0 & \cdots & 0 & a_{ik}^{(k)} & \cdots & a_{ij}^{(k)} & \cdots & a_{in}^{(k)} \\
\vdots & & \vdots & \vdots & & \vdots & & \vdots \\
0 & \cdots & 0 & a_{nk}^{(k)} & \cdots & a_{nj}^{(k)} & \cdots & a_{nn}^{(k)}
\end{array}\right]$$

我们的任务是描述怎样从 $A^{(k)}$ 得到 $A^{(k+1)}$. 为了在第 k 列的主元素 $a_{kk}^{(k)}$ 下面产生 0，我们把它下面的行减去第 k 行的倍数. 第 $i=1, 2, \cdots, k$ 行不改变. 所以公式是 [165]

$$a_{ij}^{(k+1)} = \begin{cases} a_{ij}^{(k)} & \text{若 } i \leqslant k \\ a_{ij}^{(k)} - (a_{ik}^{(k)}/a_{kk}^{(k)})a_{kj}^{(k)} & \text{若 } i \geqslant k+1 \text{ 和 } j \geqslant k+1 \\ 0 & \text{若 } i \geqslant k+1 \text{ 和 } j \leqslant k \end{cases} \tag{8}$$

然后，我们取 $U=A^{(n)}$ 并定义 L 为

$$\ell_{ik} = \begin{cases} a_{ik}^{(k)}/a_{kk}^{(k)} & \text{若 } i \geqslant k+1 \\ 1 & \text{若 } i = k \\ 0 & \text{若 } i \leqslant k-1 \end{cases} \tag{9}$$

这里 $A=LU$ 是矩阵 A 的标准高斯分解，L 为单位下三角阵，而 U 为上三角阵. 从(8)和(9)式以及前面的数值例子我们应该明白这个完整的消元过程当任何一个主元素为 0 时将中断. 现在我们证明下列定理.

定理 1（*非零主元定理*） *若在刚才描述的过程中全部主元素 $a_{kk}^{(k)}$ 非零，则 $A=LU$.*

证明 我们看到当 $i \leqslant k$ 或 $j \leqslant k-1$ 时，$a_{ij}^{(k+1)}=a_{ij}^{(k)}$. 其次注意到 $u_{kj}=a_{kj}^{(n)}=a_{kj}^{(k)}$. 最后，注意到当 $k>i$ 时，$\ell_{ik}=0$ 以及 $k>j$ 时 $u_{kj}=0$. 现在设 $i \leqslant j$. 利用这些事实，我们有

$$\begin{aligned}(LU)_{ij} &= \sum_{k=1}^{n} \ell_{ik} u_{kj} = \sum_{k=1}^{i} \ell_{ik} u_{kj}^{(k)} = \sum_{k=1}^{i} \ell_{ik} a_{kj}^{(k)} \\ &= \sum_{k=1}^{i-1} \ell_{ik} a_{kj}^{(k)} + \ell_{ii} a_{ij}^{(i)} \\ &= \sum_{k=1}^{i-1} (a_{ik}^{(k)}/a_{kk}^{(k)}) a_{kj}^{(k)} + a_{ij}^{(i)} \\ &= \sum_{k=1}^{i-1} (a_{ij}^{(k)} - a_{ij}^{(k+1)}) + a_{ij}^{(i)} \\ &= a_{ij}^{(1)} = a_{ij}\end{aligned}$$

类似地，因为 $i \geqslant j+1$ 且 $k \geqslant j+1$ 时，$a_{ij}^{(k)}=0$，所以当 $i>j$ 时，有

$$\begin{aligned}(LU)_{ij} &= \sum_{k=1}^{j} \ell_{ik} a_{kj}^{(k)} \\ &= \sum_{k=1}^{j} (a_{ij}^{(k)} - a_{ij}^{(k+1)}) \\ &= a_{ij}^{(1)} - a_{ij}^{(j+1)} \\ &= a_{ij}^{(1)} = a_{ij}\end{aligned}$$

[166] ■

对矩阵 $A=(a_{ij})$ 执行刚才所述的基本的高斯消元法的算法如下：

```
input n, (a_ij)
for k=1 to n-1 do
    for i=k+1 to n do
```

```
            z ← a_ik / a_kk
            a_ik ← 0
            for j = k+1 to n do
                a_ij ← a_ij − z a_kj
            end do
        end do
    end do
    output (a_ij)
```

这里我们假定所有主元素非零. 选择乘子使得 A 中主对角线以下的元素计算结果为 0. 与其执行这个计算，我们不如就简单地在算法中用 0 代替这些元素.

4.3.2 选主元

我们对刚才所述简单形式的高斯算法不太满意，因为它对事实上很容易求解的方程组失败. 为说明这种看法，我们考察三个基本的例子. 第一个是

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \tag{10}$$

因为无法把第 1 个方程的倍数加到第 2 个方程使第 2 个方程中 x_1 的系数为 0，所以简单形式的算法失败.（见习题 4.2.7.）

刚才遇到的困难在下列方程组中继续存在：

167

$$\begin{bmatrix} \varepsilon & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \tag{11}$$

这里 ε 是不等于 0 的小数. 当方法应用于(11)时，高斯算法产生下列上三角方程组

$$\begin{bmatrix} \varepsilon & 1 \\ 0 & 1-\varepsilon^{-1} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2-\varepsilon^{-1} \end{bmatrix} \tag{12}$$

其解是

$$\begin{cases} x_2 = (2-\varepsilon^{-1})/(1-\varepsilon^{-1}) \approx 1 \\ x_1 = (1-x_2)\varepsilon^{-1} \approx 0 \end{cases} \tag{13}$$

在计算机中，当 ε 足够小时，$2-\varepsilon^{-1}$ 将被算作和 $-\varepsilon^{-1}$ 一样大小. 同样地，分母 $1-\varepsilon^{-1}$ 将被算作和 $-\varepsilon^{-1}$ 一样大小. 因此，在这些情况下，算得的 x_2 为 1 而 x_1 为 0. 因为，正确解是

$$\begin{cases} x_1 = 1/(1-\varepsilon) \approx 1 \\ x_2 = (1-2\varepsilon)/(1-\varepsilon) \approx 1 \end{cases}$$

因此所计算的解对 x_2 是正确的，而对 x_1 是极其不正确的！

在计算机中，当 ε 足够小时，为什么 $2-\varepsilon^{-1}$ 的计算会得到像 $-\varepsilon^{-1}$ 的计算那样相同的机器数呢？理由是执行减法之前，在 2 和 ε^{-1} 的浮点形式中的指数必须通过小数点位移使之相同. 如果这个位移足够大，那么 2 的尾数就为 0. 例如，在一台类似于假想 Marc-32 的七位十进制机器上，$\varepsilon = 10^{-8}$，我们有 $\varepsilon^{-1} = 0.100\,000\,0 \times 10^9$，$2 = 0.200\,000\,0 \times 10^1$，如果用指数 9 重写，我们有 $2 = 0.000\,000\,002 \times 10^9$ 而 $2-\varepsilon^{-1} = -0.099\,999\,998 \times 10^9$，故在机器中，$2-\varepsilon^{-1} = -0.100\,000\,0 \times 10^9 = -\varepsilon^{-1}$.

最后的例子将说明引起麻烦的实际上不是系数 a_{11} 的微小．更确切地说，引起麻烦的是 a_{11} 相对于该行中其他元素太小．考察下列方程组，它等价于方程组(11)：

$$\begin{bmatrix} 1 & \varepsilon^{-1} \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \varepsilon^{-1} \\ 2 \end{bmatrix} \tag{14}$$

简单的高斯算法产生

$$\begin{bmatrix} 1 & \varepsilon^{-1} \\ 0 & 1-\varepsilon^{-1} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \varepsilon^{-1} \\ 2-\varepsilon^{-1} \end{bmatrix} \tag{15}$$

(15)的解是

$$\begin{cases} x_2 = (2-\varepsilon^{-1})/(1-\varepsilon^{-1}) \approx 1 \\ x_1 = \varepsilon^{-1} - \varepsilon^{-1}x_2 \approx 0 \end{cases}$$

对小的 ε，与前面一样还是算得 x_2 为 1 和 x_1 为 0，错误！

如果把方程的次序改变一下，那么这些例子中的困难将不会出现．因而，交换方程组(11)中的两个方程得到

168

$$\begin{bmatrix} 1 & 1 \\ \varepsilon & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

应用高斯消元法于这个方程组产生

$$\begin{bmatrix} 1 & 1 \\ 0 & 1-\varepsilon \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1-2\varepsilon \end{bmatrix}$$

所以解是

$$\begin{cases} x_2 = (1-2\varepsilon)/(1-\varepsilon) \approx 1 \\ x_1 = 2 - x_1 \approx 1 \end{cases}$$

从这些基本的例子得到的结论是当情况需要时好的算法中必须加入交换方程组中的方程．为了计算经济的理由，我们不喜欢在计算机的存储器中移动矩阵的行．代之，我们按逻辑的方式简单地选择主行．假如我们使用行 p_1，p_2，…，p_{n-1} 代替使用按 1，2，…，$n-1$ 这样次序的行作为主行．于是在第 1 步，从其他行减去第 p_1 行的倍数．若我们引进元素 p_n 使(p_1，p_2，…，p_n)是(1，2，…，n)的一个置换，则第 p_n 行将不作为主行出现，但是我们可以说从行 p_2，p_3，…，p_n 中减去第 p_1 行的倍数．在下一步中，从行 p_3，p_4，…，p_n 中减去第 p_2 行的倍数；等等．

下面是一个实现这个过程的算法．(假定置换数组 p 被预先确定并按自然数 1，2，…，n 的某个次序组成．)

```
input n, (a_ij), (p_i)
for k=1 to n-1 do
    for i=k+1 to n do
        z ← a_{p_i k} / a_{p_k k}
        a_{p_i k} ← 0
        for j=k+1 to n do
            a_{p_i j} ← a_{p_i j} - z a_{p_k j}
        end do
    end do
```

```
end do
output (a_ij)
```

把这个算法与基本的高斯消元法相比较，我们看到除了一个整体的交换外它们是恒等的. 在上面的伪代码中，系数数组 A 元素的首下标涉及置换数组 p. 当然，当置换数组对应于自然次序($p_i=i$)时，将得到基本的方法.

4.3.3 行尺度主元高斯消元法

169

现在我们描述求解 $n\times n$ 方程组

$$Ax = b$$

的一个称为**行尺度主元高斯消元法**的算法. 算法由两部分组成：**分解阶段**(也称为**向前消元**)和**求解阶段**(包含**更新**和**向后回代**). 分解阶段只应用于 A 并设计产生 PA 的 LU 分解，这里 P 是从置换数组 p 导出的一个置换矩阵. (PA 是通过对 A 行置换得到的矩阵.)置换后的线性方程组是

$$PAx = Pb$$

分解 $PA=LU$ 是从下面要说明的修正的高斯消元法得到的. 在求解阶段，我们考虑两个方程 $Lz=Pb$ 和 $Ux=z$. 首先，右端项 b 按 P 重排并且将所得的结果放回 b 中；即 $b\leftarrow Pb$. 其次，从 $Lz=b$ 解出 z 并且把所得结果放回到 b 中；即 $b\leftarrow L^{-1}b$. 因为 L 是单位下三角阵，这等于向前回代过程. 这个过程称为**更新** b. 然后再用回代过程求解 $Ux=b$ 得到 x_n，x_{n-1}，…，x_1.

在分解阶段，我们先计算每行的**尺度**. 取

$$s_i=\max_{1\leqslant j\leqslant n}|a_{ij}|=\max\{|a_{i1}|,\ |a_{i2}|,\ \cdots,\ |a_{in}|\}\qquad(1\leqslant i\leqslant n)$$

这些值被记录在算法的一个数组 s 中.

在分解阶段开始时，我们不从其他行中任意减去第 1 行的倍数. 而是选择使 $|a_{i1}|/s_i$ 最大的行为**主行**. 这样选择的指标用 p_1 表示并且变成置换数组中的第 1 个元素. 于是，对 $1\leqslant i\leqslant n$，$|a_{p_1 1}|/s_{p_1}\geqslant|a_{i1}|/s_i$. 一旦确定 p_1，为了在 A 的第 1 列中产生 0，我们从其他行中减去第 p_1 行的适当倍数. 当然，在整个分解过程的剩余部分第 p_1 行保持不变.

为保留所产生指标 p_i 的轨迹，开始我们取置换向量(p_1，p_2，…，p_n)为(1，2，…，n). 然后，选择使 $|a_{p_j 1}|/s_{p_j}$ 为最大的指标 j 并且在置换数组 p 中用 p_j 交换 p_1. 实际的消元步骤中包含从第 p_i 行减去第 p_1 行的($a_{p_i 1}/a_{p_1 1}$)倍，其中 $2\leqslant i\leqslant n$.

为描述一般的过程，假如我们已经在第 k 列中产生 0. 我们对 $k\leqslant i\leqslant n$，审视数 $|a_{p_i k}|/s_{p_i}$，求最大元. 若 j 是这些比中第 1 个最大的指标，则我们在数组 p 中用 p_j 交换 p_k，然后，对 $k+1\leqslant i\leqslant n$，从第 p_i 行减去第 p_k 行的($a_{p_i k}/a_{p_k k}$)倍.

下面说明对矩阵

$$A=\begin{bmatrix}2 & 3 & -6\\ 1 & -6 & 8\\ 3 & -2 & 1\end{bmatrix}$$

如何做这项工作. 一开始，$p=(1,2,3)$，$s=(6,8,3)$. 为选择第 1 个主行，看比{2/6，1/8，3/3}. 最大的比对应 $j=3$，而第 3 行是第 1 个主行. 故我们用 p_3 交换 p_1 得到 $p=(3,2,1)$. 现

在从第 2 行和第 1 行中减去第 3 行的倍数并在第 1 列中产生 0. 结果是 [170]

$$\begin{bmatrix} \textcircled{\frac{2}{3}} & \frac{13}{3} & -\frac{20}{3} \\ \textcircled{\frac{1}{3}} & -\frac{16}{3} & \frac{23}{3} \\ 3 & -2 & 1 \end{bmatrix}$$

在 a_{11} 和 a_{21} 位置上画圈的元素是乘子. 在下一步，主行的选择是根据数 $|a_{p_2 2}|/s_{p_2}$ 和 $|a_{p_3 2}|/s_{p_3}$ 作出的. 这些比中第 1 个是(16/3)/8，第 2 个是(13/3)/6. 故 $j=3$，我们用 p_3 交换 p_2. 然后从第 p_3 行中减去第 p_2 行的倍数. 结果是 $p=(3,1,2)$ 和

$$\begin{bmatrix} \textcircled{\frac{2}{3}} & \frac{13}{3} & -\frac{20}{3} \\ \textcircled{\frac{1}{3}} & \textcircled{-\frac{16}{13}} & -\frac{7}{13} \\ 3 & -2 & 1 \end{bmatrix}$$

最后的乘子存放在 a_{22} 位置上.

若原来矩阵 A 的行已经按最后的置换数组 p 作了交换，则我们将有 A 的 LU 分解. 因此，有

$$PA = \begin{bmatrix} 1 & 0 & 0 \\ \frac{2}{3} & 1 & 0 \\ \frac{1}{3} & -\frac{16}{13} & 1 \end{bmatrix} \begin{bmatrix} 3 & -2 & 1 \\ 0 & \frac{13}{3} & -\frac{20}{3} \\ 0 & 0 & -\frac{7}{13} \end{bmatrix} = \begin{bmatrix} 3 & -2 & 1 \\ 2 & 3 & -6 \\ 1 & -6 & 8 \end{bmatrix}$$

其中

$$P = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \qquad A = \begin{bmatrix} 2 & 3 & -6 \\ 1 & -6 & 8 \\ 3 & -2 & 1 \end{bmatrix}$$

置换阵 P 是由置换数组 p 取 $(P)_{ij}=\delta_{p_i j}$ 而得到的. 换言之，P 按 p 中元素来置换单位阵 I 的行形成的.

下面是执行行尺度主元高斯消元法**分解阶段**的一个算法：

```
input n, (a_ij)
for i=1 to n do
    p_i ← i
    s_i ← max_{1≤j≤n} |a_ij|
end do
for k=1 to n−1 do
    select j≥k so that
        |a_{p_j k}|/s_{p_j} ≥ |a_{p_i k}|/s_{p_i} for i=k, k+1, …, n
    p_k ↔ p_j
    for i=k+1 to n do
```

[171]

```
z ← a_{p_i k}/a_{p_k k}; a_{p_i k} ← z
for j=k+1 to n do
    a_{p_i j} ← a_{p_i j} − z a_{p_k j}
end do
end do
end do
output (a_ij), (p_i)
```

注意，**乘子**被存放在 A 中那些由消元过程产生 0 的位置上．因此，所有为重构 LU 分解的必要数据被存放在这个数组中．虽然如此，用户应该意识到算法覆盖了原来 A 数组中的值．如果再次需要矩阵 A 的话，应该把它存储在另一个数组中．

一旦对 A 已经执行了分解阶段，为了求解 $Ax=b$，我们利用最后的置换数组 p 以及算法分解阶段已经确定的乘子对 b 应用算法的向前阶段．接下去，我们执行向后回代；即按次序求解 x_n，x_{n-1}，…，x_1．下面是执行算法**求解阶段**的算法：

```
input n, (a_ij), (p_i), (b_i)
for k=1 to n-1 do
    for i=k+1 to n do
        b_{p_i} ← b_{p_i} − a_{p_i k} b_{p_k}
    end do
end do
for i=n to 1 step −1 do
    x_i ← (b_{p_i} − Σ_{j=i+1}^{n} a_{p_i j} x_j) / a_{p_i i}
end do
output (x_i)
```

172

伪代码的第一部分对应于分解阶段期间发生的情况更新右边的 b．第二部分从下三角方程组中求 x，即向后回代．当然，置换数组 p 必须防止被搞乱．注意，当 $i=n$ 时 j-和是空的，故 $x_n=b_{p_n}/a_{p_n n}$．

4.3.4　全主元高斯消元法

还有另一种选主元方法是**全主元法**．记得在部分选主元中，通过考察子列下面包括对角元 $a_{kk}^{(k-1)}$ 的 $n-k+1$ 个元素来确定第 k 个主元素．在部分选主元时，把这些元素中绝对值最大的元素所在的行命名为主行，在尺度部分选主元中，形成这些元素同行尺度常数之比，并且把这些比的绝对值最大的相应行选为主行．反之，全主元类似于部分选主元，要求考察右下方子阵，包括对角元 $a_{kk}^{(k-1)}$ 在内的所有 $(n-k+1)^2$ 个元素(同样可定义尺度全主元．) 一般说来，因为额外的计算量，(选择主元素需要考察矩阵中许许多多的元素) 所以认为越过部分主元来选择全主元是没有实用价值的．

4.3.5　分解 $PA=LU$

如前面所提到的那样，若行尺度主元包括在高斯算法中，则我们得到 PA 的 LU 分解，这里 P 是一个确定的置换阵．现在仿效定理 1 的证明给出这个结果的一个证明．该证明与所使用的确定主元的方法不相关．

设 p_1，p_2，…，p_n 是它们变成主行的行序指标. 设 $A^{(1)}=A$，用公式

$$a_{p_ij}^{(k+1)}=\begin{cases}a_{p_ij}^{(k)} & \text{若 } i\leqslant k \text{ 或 } i>k>j\\ a_{p_ij}^{(k)}-(a_{p_ik}^{(k)}/a_{p_kk}^{(k)})a_{p_kj}^{(k)} & \text{若 } i>k \text{ 和 } j>k\\ a_{p_ik}^{(k)}/a_{p_kk}^{(k)} & \text{若 } i>k \text{ 和 } j=k\end{cases} \tag{16}$$

递归地定义 $A^{(2)}$，$A^{(3)}$，…，$A^{(n)}$.

定理 2（PA 是 LU 分解定理） 定义置换阵 P，它的元素是 $P_{ij}=\delta_{p_ij}$. 定义上三角阵 U，它的元素是 $u_{ij}=a_{p_ij}^{(n)}$，$j\geqslant i$. 定义单位下三角阵 L，它的元素是 $\ell_{ij}=a_{p_ij}^{(n)}$，$j<i$. 于是，$PA=LU$.

证明 从递归公式，我们有

$$u_{kj}=a_{p_kj}^{(n)}=a_{p_kj}^{(k)}\qquad(j\geqslant k)$$

$$\ell_{ik}=a_{p_ik}^{(n)}=a_{p_ik}^{(k+1)}=a_{p_ik}^{(k)}/a_{p_kk}^{(k)}\qquad(i\geqslant k)$$

这两个等式与下列事实有关：$A^{(n)}$ 中的第 p_k 行在第 k 步变成固定的，而 $A^{(n)}$ 中的第 k 列在第 $k+1$ 步变成固定的. 于是，

$$a_{p_kj}^{(n)}=a_{p_kj}^{(k)} \text{ 和 } a_{p_ik}^{(n)}=a_{p_ik}^{(k+1)}$$

173

进而，当 $i=k$ 时，因为公式对 ℓ_{kk} 产生 1，所以刚才给出的 ℓ_{ik} 公式是正确的. 现在假如 $i\leqslant j$，则

$$\begin{aligned}(LU)_{ij}&=\sum_{k=1}^{i}\ell_{ik}u_{kj}\\&=\sum_{k=1}^{i-1}(a_{p_ik}^{(k)}/a_{p_kk}^{(k)})a_{p_kj}^{(k)}+\ell_{ii}a_{p_ij}^{(i)}\\&=\sum_{k=1}^{i-1}(a_{p_ij}^{(k)}-a_{p_ij}^{(k+1)})+a_{p_ij}^{(i)}\\&=a_{p_ij}^{(1)}=a_{p_ij}\end{aligned}$$

若 $i>j$，则

$$\begin{aligned}(LU)_{ij}&=\sum_{k=1}^{j}\ell_{ik}u_{kj}\\&=\sum_{k=1}^{j-1}(a_{p_ik}^{(k)}/a_{p_kk}^{(k)})a_{p_kj}^{(k)}+(a_{p_ij}^{(j)}/a_{p_jj}^{(j)})a_{p_jj}^{(j)}\\&=\sum_{k=1}^{j-1}(a_{p_ij}^{(k)}-a_{p_ij}^{(k+1)})+a_{p_ij}^{(j)}\\&=a_{p_ij}^{(1)}=a_{p_ij}\end{aligned}$$

另一方面，

$$(PA)_{ij}=\sum_{k=1}^{n}P_{ik}a_{kj}=\sum_{k=1}^{n}\delta_{p_ik}a_{kj}=a_{p_ij}$$

因此，我们已证明，对一切(i,j)，

$$(PA)_{ij}=(LU)_{ij}$$ ■

定理 3（求解 $PA=LU$ 的定理） 若分解 $PA=LU$ 是由行尺度主元高斯算法产生的，则 $Ax=b$ 的解通过先解 $Lz=Pb$，再解 $Ux=z$ 得到. 类似地，$y^{\mathrm{T}}A=c^{\mathrm{T}}$ 的解通过先解 $U^{\mathrm{T}}z=c$ 再解 $L^{\mathrm{T}}Py=z$ 得到.

证明 留作习题 4.3.47. ■

[174] 根据 L 和 U 求解 $Ax=b$ 的伪代码如下：

input n，(ℓ_{ij})，(u_{ij})，(b_i)，(p_i)
for $i=1$ **to** n **do**

$$z_i \leftarrow b_{p_i} - \sum_{j=1}^{i-1} \ell_{ij} z_j$$

end do
for $i=n$ **to** 1 **step** -1 **do**

$$x_i \leftarrow \left(z_i - \sum_{j=i+1}^{n} u_{ij} x_j\right) \Big/ u_{ii}$$

end do
output (x_i)

可用由高斯消元法得到的最后的 A 矩阵元素编写同样的算法．结果是：

input n，(a_{ij})，(b_i)，(p_i)
for $i=1$ **to** n **do**

$$z_i \leftarrow b_{p_i} - \sum_{j=1}^{i-1} a_{p_i j} z_j$$

end do
for $i=n$ **to** 1 **step** -1 **do**

$$x_i \leftarrow \left(z_i - \sum_{j=i+1}^{n} a_{p_i j} x_j\right) \Big/ a_{p_i i}$$

end do
output (x_i)

求解 $y^{\mathrm{T}}A=c^{\mathrm{T}}$ 的伪代码如下(记住 $P^{-1}=P^{\mathrm{T}}$)：

input n，(a_{ij})，(c_i)，(p_i)
for $j=1$ **to** n **do**

$$z_j \leftarrow \left(c_j - \sum_{i=1}^{j-1} a_{p_i j} z_i\right) \Big/ a_{p_j j}$$

end do
for $j=n$ **to** 1 **step** -1 **do**

$$y_{p_j} \leftarrow z_j - \sum_{i=j+1}^{n} a_{p_i j} y_{p_i}$$

end do
output (y_i)

4.3.6 运算量

为了估计求解线性方程组的计算效果，我们应该计算分解阶段和求解阶段的算术运算次数．因为乘法和除法运算执行的时间通常是相当的并且比加法和减法耗时多得多．所以传统上只计算乘法和除法的次数，把它们归并在一起作为**长运算**或简称 **op**. [175]

考察分解过程中的第 1 个主步($k=1$)，它对 $n\times n$ 矩阵 A 运算．必须计算确定 p_1(主行的指标)包括 n 次除法(n 次 op)．对在 $n-1$ 个编号为 p_2，p_3，…，p_n 的行中的每一行，计算乘子(1 次 op)，然后从第 p_i 行($2\leqslant i\leqslant n$)减去第 p_1 行的倍数．在第 1 列中产生的 0 是不计算的．因此，乘子和每行的消元过程耗费 n 次 op．因为要以这个方式处理 $n-1$ 行，所以我们有 $n(n-1)$

次 op. 做这项工作再加上确定 p_1需要的 n 次 op. 共计 n^2 次 op.

分解过程的剩余部分可解释为对越来越小的矩阵重复第 1 步. 因此，在第 2 步中，第 p_1 行和第 1 列不再考虑. 在第 2 步中整个计算就像第 1 步那样，只不过应用于一个$(n-1)\times(n-1)$矩阵. 通过观察推出分解需要

$$n^2+(n-1)^2+\cdots+3^2+2^2=\frac{1}{3}n^3+\frac{1}{2}n^2+\frac{1}{6}n-1\approx\frac{1}{3}n^3+\frac{1}{2}n^2$$

次长运算. 这里我们利用事实

$$\sum_{k=1}^{n}k^2=\frac{1}{6}n(n+1)(2n+1)$$

对大的 n，$n^3/3$ 项是主项. 因此，如用行尺度主元求 LU 分解，对大的 n，大约含有 $n^3/3$ 次长运算.

检查算法的第 2 阶段表明，在更新右端项 b 中，有 $n-1$ 步，在其第 1 步，有 $n-1$ 次长运算，在第 2 步有 $n-2$ 次长运算，等等. 所以总计

$$(n-1)+(n-2)+\cdots+1=\frac{1}{2}n^2-\frac{1}{2}n$$

这里，我们使用

$$\sum_{k=1}^{n}k=\frac{1}{2}n(n+1)$$

在回代中，第 1 步(计算 x_n)有一次长运算. 然后依次有 2，…，n 次长运算. 总计是

$$1+2+3+\cdots+n=\frac{1}{2}n^2+\frac{1}{2}n$$

所以算法的这个阶段最大的总数是 n^2. 总结这些结果，下面我们叙述更加一般的结果.

定理 4(长运算定理) *若高斯消元法用行尺度主元，则对固定的 A 和 m 个不同的向量 b，方程组 $Ax=b$ 的解大致含有*

$$\frac{1}{3}n^3+\left(\frac{1}{2}+m\right)n^2$$

次长运算(乘法和除法).

176

高斯消元法的结构允许有效地处理问题 $Ax^{(i)}=b^{(i)}$，$i=1, 2, \cdots, m$. 这里我们有 m 个线性方程，其中每个方程具有相同的系数阵但是有不同的右端项. 我们应用向前消去阶段给出分解 $PA=LU$. 然后为了得到 $x^{(i)}$，需要利用这个分解作 m 次"回代求解". 因此，只需要 $\mathcal{O}(n^3/3+(1/2+m)n^2)$次长运算而不是分别处理 m 个方程组时所需要的 $\mathcal{O}(mn^3/3)$次长运算. 进而，我们可以用 $\mathcal{O}(4/3n^3)$次 op 计算 A^{-1}，因为它是我们前面所讨论情况的特例；即 $AX=I$ 可以对 n 个 i 的值用 $Ax^{(i)}=e_i$ 求解 A^{-1}的每列. 忠告读者，当求解 $Ax=b$ 时不要计算 A^{-1}而应该直接求解 x！

4.3.7 对角占优矩阵

有时一个方程组具有性质：不选主元的高斯消元法可以安全地用于求解. 具有上述性质的一类矩阵是**对角占优矩阵**. 这个性质用不等式表达为

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \qquad (1 \leqslant i \leqslant n) \tag{17}$$

下面是一个对角占优矩阵的例子：

$$\begin{bmatrix} 4 & -1 & 0 & -1 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix}$$

这种矩阵自然地出现在涉及偏微分方程用有限差分离散化的应用中，它们还出现在样条和许多其他领域的研究中.

若系数阵具有这个性质，则在高斯消元法的第 1 步中，因为由不等式(17)知主元素 a_{11} 不为 0，所以我们可用第 1 行作为主行. 在完成第 1 步后，我们知道第 2 行就可用作下一个主行. 这种情况由下面的定理决定.

定理 5(*保持对角占优定理*) *不选主元高斯消元法保持矩阵的对角占优性.*

证明 考察高斯消元法的第 1 步就足够了(第 1 列在该步中产生 0)，因为后续步除了应用于更小的矩阵外非常像第 1 步. 因此，设 A 是一个 $n\times n$ 对角占优阵. 考虑到在第 1 列中产生了 0 以及第 1 行不变的事实，我们必须对 $i=1, 2, \cdots, n$ 证明

177

$$|a_{ii}^{(2)}| > \sum_{\substack{j=2 \\ j \neq i}}^{n} |a_{ij}^{(2)}|$$

依据 A，这意味着

$$|a_{ii} - (a_{i1}/a_{11})a_{1i}| > \sum_{\substack{j=2 \\ j \neq i}}^{n} |a_{ij} - (a_{i1}/a_{11})a_{1j}|$$

足够证明更强的不等式.

$$|a_{ii}| - |(a_{i1}/a_{11})a_{1i}| > \sum_{\substack{j=2 \\ j \neq i}}^{n} \{|a_{ij}| + |(a_{i1}/a_{11})a_{1j}|\}$$

一个等价的不等式是

$$|a_{ii}| - \sum_{\substack{j=2 \\ j \neq i}}^{n} |a_{ij}| > \sum_{j=2}^{n} |(a_{i1}/a_{11})a_{1j}|$$

从第 i 行对角占优性，我们知道

$$|a_{ii}| - \sum_{\substack{j=2 \\ j \neq i}}^{n} |a_{ij}| > |a_{i1}|$$

因此，只需证明下式即可.

$$|a_{i1}| \geqslant \sum_{j=2}^{n} |(a_{i1}/a_{11})a_{1j}|$$

因为第 1 行的对角占优性，

$$|a_{11}| > \sum_{j=2}^{n} |a_{1j}| \Longrightarrow 1 > \sum_{j=2}^{n} |a_{1j}/a_{11}|$$

所以前式成立. ■

推论 1(对角占优矩阵第一推论) *每个对角占优矩阵非奇异且有 LU 分解.*

证明 定理 5 连同定理 1 推出对角占优矩阵 A 有 LU 分解，其中 L 是单位下三角阵. 由前面的定理知道，矩阵 U 对角占优. 因此它的对角元非零，所以 L 和 U 是非奇异的. ■ [178]

推论 2(对角占优矩阵第二推论) *若高斯消元法的行尺度主元形式在每个主步以后重新计算尺度数组并应用于对角占优阵，则主行将是自然次序：1，2，…，n. 所以，此时选主元的工作可以略去.*

证明 由定理 5，只需证明算法中选择的第 1 主元是 1 就足够了. 因此，只需证明

$$|a_{11}|/s_1>|a_{i1}|/s_i \qquad (2\leqslant i\leqslant n)$$

即可. 根据对角占优性，对一切 i，$|a_{ii}|=\max_j|a_{ij}|=s_i$. 因此，$|a_{11}|/s_1=1$. 对 $i\geqslant 2$，我们有

$$|a_{i1}|\leqslant \sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|<|a_{ii}|=s_i$$

于是，$|a_{i1}|/s_i<1$. ■

在 4.2 节中已证明了对称正定阵 A 有唯一的楚列斯基分解 $A=LL^{\mathrm{T}}$.（4.2 节中的定理 2.）也证明了下三角阵 L 的元素满足不等式

$$|\ell_{ij}|\leqslant\sqrt{a_{ii}} \tag{18}$$

楚列斯基分解算法是 LU 分解的特例. 因为如不等式(18)所指出的那样，L 的元素与 A 的元素相比数量上还是适度的，所以不需要选主元.

4.3.8 三对角方程组

在应用中，方程组时常出现有特殊结构的系数矩阵. 利用特殊结构的特制算法来求解这些方程组通常较好. 我们考虑其中的一个例子是三对角方程组.

如果对一切满足 $|i-j|>1$ 的对(i,j)，$a_{ij}=0$，那么方阵 $A=(a_{ij})$称为**三对角**. 因此，在第 i 行只有三个不为 0 的元素 $a_{i,i-1}$，a_{ii}，$a_{i,i+1}$. 有三个向量可用于存放非零元，我们安排记号如下：

$$\begin{bmatrix} d_1 & c_1 & & & & & \\ a_1 & d_2 & c_2 & & & & \\ & a_2 & d_3 & c_3 & & & \\ & & a_3 & d_4 & c_4 & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & a_{n-2} & d_{n-1} & c_{n-1} \\ & & & & & a_{n-1} & d_n \end{bmatrix}\begin{bmatrix} x_1\\ x_2\\ x_3\\ x_4\\ \vdots\\ x_{n-1}\\ x_n\end{bmatrix}=\begin{bmatrix} b_1\\ b_2\\ b_3\\ b_4\\ \vdots\\ b_{n-1}\\ b_n\end{bmatrix} \tag{19}$$

在上式中未标出的矩阵元素都为 0. [179]

我们将假设系数矩阵在求解方程组中不需要选主元. 例如，当矩阵对称正定时这是成立的. 将使用简单的高斯消元法，并且在此算法中同时处理右端项(向量 b). 在第 1 步，第 2 行

减去第 1 行的倍数产生一个 0，其中 a_1 维持原状．注意 d_2 和 b_2 是改变的但 c_2 不变．适当的倍数是 a_1/d_1．因此第 1 步由下列替换组成：

$$d_2 \leftarrow d_2 - (a_1/d_1)c_1$$
$$b_2 \leftarrow b_2 - (a_1/d_1)b_1$$

所有向前消元阶段的剩余步完全像第 1 步．在回代阶段，第 1 步是

$$x_n \leftarrow b_n/d_n$$

下一步是

$$x_{n-1} \leftarrow (b_{n-1} - c_{n-1}x_n)/d_{n-1}$$

所有其余步都是类似的．下面是算法 tri：

```
input n, (a_i), (b_i), (c_i), (d_i)
for i=2 to n do
    d_i ← d_i − (a_{i−1}/d_{i−1})c_{i−1}
    b_i ← b_i − (a_{i−1}/d_{i−1})b_{i−1}
end do
x_n ← b_n/d_n
for i=n−1 to 1 step −1 do
    x_i ← (b_i − c_i x_{i+1})/d_i
end do
output (x_i)
```

习题 4.3

1. 两次求解下列线性方程组．第一次使用高斯消元法并给出分解 $A=LU$．第二次使用行尺度主元高斯消元法并确定分解 $PA=LU$．

a. $\begin{bmatrix} -1 & 1 & -4 \\ 2 & 2 & 0 \\ 3 & 3 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}=\begin{bmatrix} 0 \\ 1 \\ \frac{1}{2} \end{bmatrix}$

b. $\begin{bmatrix} 1 & 6 & 0 \\ 2 & 1 & 0 \\ 0 & 2 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}=\begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix}$

[180]

c. $\begin{bmatrix} -1 & 1 & 0 & -3 \\ 1 & 0 & 3 & 1 \\ 0 & 1 & -1 & -1 \\ 3 & 0 & 1 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}=\begin{bmatrix} 4 \\ 0 \\ 3 \\ 1 \end{bmatrix}$

d. $\begin{bmatrix} 6 & -2 & 2 & 4 \\ 12 & -8 & 4 & 10 \\ 3 & -13 & 3 & 3 \\ -6 & 4 & 2 & -18 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}=\begin{bmatrix} 0 \\ -10 \\ -39 \\ -16 \end{bmatrix}$

e. $\begin{bmatrix} 1 & 0 & 2 & 1 \\ 4 & -9 & 2 & 1 \\ 8 & 16 & 6 & 5 \\ 2 & 3 & 2 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}=\begin{bmatrix} 2 \\ 14 \\ -3 \\ 0 \end{bmatrix}$

2. 说明确定高斯消元法的(8)式也能写成下面形式

$$a_{ij}^{(k+1)} = \begin{cases} a_{ij}^{(k)} & \text{若 } i \leqslant k \text{ 或 } j < k \\ a_{ij}^{(k)} - (a_{ik}^{(k)}/a_{kk}^{(k)})a_{kj}^{(k)} & \text{若 } i > k \text{ 且 } j \geqslant k \end{cases}$$

3. 设$(p_1, p_2, \cdots, p_n)$是$(1, 2, \cdots, n)$的一个置换并由$P_{ij}=\delta_{p_i j}$定义矩阵P. 设A是任意$n\times n$矩阵. 描述PA, AP, P^{-1}, PAP^{-1}.
4. 设A是具有尺度因子$s_i=\max\limits_{1\leqslant j\leqslant n}|a_{ij}|$的$n\times n$矩阵. 假定所有$s_i$是正的，并设$B$是一个元素为$(a_{ij}/s_i)$的矩阵. 证明：若向前消元法应用于$A$和$B$，则它们两个最后的$L$数组相同. 求有关最后的$A$和$B$矩阵(处理后的)公式.
5. 引入新变量$y_i=d_ix_i$，这里d_i是正数，修正方程组$Ax=b$有时是明智的. 若x对应于物理量，则这个变量的改变对应于度量x_i的单位的改变. 因此，若我们决定改变x_1从分米到米，则$y_1=10^{-2}x_1$. 按矩阵的术语，我们定义一个具有对角元d_i的对角阵D，并取$y=Dx$. 新的方程组是$AD^{-1}y=b$. 若d_j选为$\max\limits_{1\leqslant i\leqslant n}|a_{ij}|$，我们称之为**列均衡化**. 结合列均衡化修正分解和求解算法.（两个算法一起仍将求解$Ax=b$.）
6. 说明全主元(行和列同时选主元)高斯算法中的乘子位于区间$[-1, 1]$中.（见计算机习题4.3.1.）
7. 设$n\times n$矩阵A用向前消元法处理，所得的矩阵称为B，置换向量$p=(p_1, p_2, \cdots, p_n)$. 设$P$是把单位阵的行按次序$p_1, p_2, \cdots, p_n$写出后得到的矩阵. 证明如下可得到$PA$的$LU$分解：取$C=PB$，$L_{ij}=C_{ij}$，$j<i$而$U_{ij}=C_{ij}$，$i\leqslant j$.（当$i>j$时$U_{ij}=0$，当$j>i$时$L_{ij}=0$并且$L_{ii}=1$.）
8. 当已知A的LU分解中因子U时，计算L的算法是怎样的?
9. 说明如何对本例

$$\begin{bmatrix} 2 & -2 & -4 \\ 1 & 1 & -1 \\ 3 & 7 & 5 \end{bmatrix}$$

用行尺度高斯消元法工作(仅仅向前阶段). 展示尺度数组(s_1, s_2, s_3)和最后的置换数组(p_1, p_2, p_3). 说明最后的A数组具有存放在正确位置上的乘子.

10. (续)对矩阵

$$\begin{bmatrix} 3 & 7 & 3 \\ 1 & \frac{7}{3} & 4 \\ 4 & \frac{4}{3} & 0 \end{bmatrix}$$

执行上述问题中的命令.

181

11. 假定A是三对角阵. 定义$c_0=0$和$a_n=0$. 说明：若A是按**列对角占优**的，

$$|d_i| > |a_i| + |c_{i-1}| \qquad (1\leqslant i\leqslant n)$$

则三对角方程组的算法理论上会成功的，因为不会遇到0主元. 记号参照(19)式.

12. 设A有性质$a_{ij}=0$，$i>j+1$，编写一个求解线性方程组$Ax=b$的特别的高斯消元法. 不用选主元. 算法中要包括右端项的处理. 计算求解$Ax=b$的次数.
13. 对三对角方程组计算课本中算法的运算次数.
14. 重写三对角方程组的算法使得处理方程和变量的次序是反向的.
15. 证明高斯消元法中有关长运算次数的定理.
16. 说明如何对本例

$$A = \begin{bmatrix} -9 & 1 & 17 \\ 3 & 2 & -1 \\ 6 & 8 & 1 \end{bmatrix}$$

用行尺度主元的高斯消元法. 指出尺度数组. 最后的 A 数组应该包含存放在正确位置上的乘子. 确定 P, L, U, 并验证 $PA=LU$.

17. 说明如何对矩阵

$$\begin{bmatrix} 1 & -2 & 3 \\ 2 & -4 & 2 \\ 3 & -5 & -1 \end{bmatrix}$$

做行尺度主元高斯消元法的分解阶段. 说明所有中间步(乘子, 尺度数组 s 和指标数组 p)以及算法完成后出现的最后数组 A.

18. 本题说明一个方程组的解相对于数据中的扰动可能不稳定. 对下列矩阵

$$A_1 = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \qquad A_2 = \begin{bmatrix} 1 & 1 \\ 1 & 0.01 \end{bmatrix}$$

中的每一个求解 $Ax=b$, 其中 $b=(100, 1)^T$. (见 Stoer and Bulirsch[1980, 第 185 页].)

19. 假定 $0<\varepsilon<2^{-22}$. 若在 Marc-32 上用不选主元的高斯算法求解方程组

$$\begin{cases} \varepsilon x_1 + 2x_2 = 4 \\ x_1 - x_2 = -1 \end{cases}$$

试问解向量(x_1, x_2)是多少?

20. 用全主元(如计算机习题 4.3.1 中所描述的)高斯消元法求解方程组

182

$$\begin{bmatrix} -9 & 1 & 17 \\ 3 & 2 & -1 \\ 6 & 8 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ 9 \\ -3 \end{bmatrix}$$

21. 用下列高斯消元法求解方程组

$$\begin{cases} 0.2641x_1 + 0.1735x_2 + 0.8642x_3 = -0.7521 \\ 0.9411x_1 + 0.0175x_2 + 0.1463x_3 = 0.6310 \\ -0.8641x_1 - 0.4243x_2 + 0.0711x_3 = 0.2501 \end{cases}$$

a. 不选主元

b. 行尺度主元

22. 在下列条件下编写一个求解方程组 $Ax=b$ 的算法: 存在$(1, 2, \cdots, n)$的一个置换$(p_1, p_2, \cdots, p_n)$使得对每个 i, 方程 p_i 只含变量 x_i.

23. (续)重复上题. 假定对每个 i, 方程 i 只含变量 x_{p_i}.

24. (续)重复上面的习题 4.3.22, 假定对每个 i, 方程 p_i 只含变量 x_{p_i}, 此时, 给出最简单算法.

25. a. 说明当我们对对称阵 A 应用不选主元的高斯消元法时, $\ell_{i1}=a_{1i}/a_{11}$.

b. 由此说明当移去 $A^{(2)}$ 的第 1 行和第 1 列时, 剩余的$(n-1)\times(n-1)$矩阵是对称的. 因而得出这个较小的矩阵对角线下面的元素不需要计算. 利用归纳法推断这个简化将在分解阶段的每个后继步中出现.

c. 说明所需要的计算几乎是非对称情况的一半.

d. 利用这个简化求解方程组

$$\begin{cases} 0.6428x_1 + 0.3475x_2 - 0.8468x_3 = 0.4127 \\ 0.3475x_1 + 1.8423x_2 + 0.4759x_3 = 1.7321 \\ -0.8468x_1 + 0.4759x_2 + 1.2147x_3 = -0.8621 \end{cases}$$

26. 考察矩阵

$$\begin{bmatrix} 0 & 4 & 25 & 79 \\ 9 & 7 & 39 & 89 \\ 0 & 16 & 2 & 99 \\ 0 & 6 & 6 & 49 \end{bmatrix}$$

把用作行尺度主元高斯消元法中下一个主元素的元素圈起来．尺度数组是 $s=(80, 89, 160, 30)$.

27. 指出对矩阵

$$\begin{bmatrix} 2 & -2 & -4 \\ 1 & 1 & -1 \\ 3 & 7 & 5 \end{bmatrix}$$

应用行尺度主元高斯消元法向前阶段以后所得的矩阵．在最后的矩阵中，把乘子写在适当的位置上.

28. 不用子式展开计算行列式．确定 $\det(A)$，其中

$$A = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 2 \end{bmatrix}$$

183

29. 利用行尺度主元高斯消元法求

$$A = \begin{bmatrix} 0 & -1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 2 & 0 \\ 2 & 0 & 1 & 0 \end{bmatrix}$$

的行列式.

30. 考察方程组

$$\begin{cases} x_2 + 2x_3 = 1 \\ 2x_1 - x_2 = 2 \\ 2x_2 + x_3 = 3 \end{cases}$$

求分解 $PA=LU$，这里 P 是置换阵．并利用这个分解得到 $\det(A)$.

31. 考察

$$A = \begin{bmatrix} 3 & 2 & -1 \\ 6 & 6 & 2 \\ -1 & 1 & 3 \end{bmatrix}$$

利用行尺度主元高斯消元法得到分解

$$PA = LDU$$

这里 L 是单位下三角阵，U 是单位上三角阵，D 是对角阵，而 P 是置换阵.

32. 在下面几个问题中，我们固定 n 并用 J 表示集合 $\{1, 2, \cdots, n\}$．J 的**置换**是一个映射 $p: J \twoheadrightarrow J$，这里双箭头表示 p 是**满射**．因此，J 的每个元素是 J 中某个元素 i 的象 $p(i)$．**恒等置换**用 $u(i)=i$（对一切 $i\in J$）来定义．通常用等式 $p \circ q(i)=p(q(i))$ 来定义 $p \circ q$. 说明若 p 和 q 是 J 的置换，则 $p \circ q$ 也是 J 的置换．证明 $p \circ (q \circ r)=(p \circ q) \circ r$ 以及 $p \circ u=u \circ p=p$.

33.（续）证明：每个置换 p 有一个逆 p^{-1} 满足 $p \circ p^{-1}=u=p^{-1} \circ p$．$J$ 的所有置换的集合称为 J 上的**对称群**.

34.（续）建立一个求任何给定置换之逆的算法．（在计算机中，J 的置换可表示为一个向量 $(p(1), p(2), \cdots, p(n))$.）

35. 我们给出一个方程组 $Ax=b$，其中 A 是 $n\times n$ 矩阵．设 p 和 q 是 $\{1, 2, \cdots, n\}$ 的置换．假设对 $i=1, 2, \cdots, n$，编号为 p_i 的方程只含变量 x_{q_i}，编写一个求解方程组的算法.

36.（续）假设对每个 i，变量 x_{q_1}，x_{q_2}，…，$x_{q_{i-1}}$，不出现在编号为 p_i 的方程中，重复处理上题.

37.（续）假设对每个 i，变量 x_{q_i} 只出现在编号为 p_1，p_2，…，p_i 的方程中．重复处理上面习题 4.3.35.

38.（难题）假设除 $|i-j|\leqslant 1$ 或 $(i,j)=(1,n)$ 或 $(i,j)=(n,1)$ 之外，所有元素 a_{ij} 全为 0．找一个求解 $Ax=b$ 的算法．利用不选主元的高斯消元法．

184

39. 假设不做选主元，计算 $n\times n$ 矩阵的 LU 分解中所包含的长运算次数．

40. 设 A 是按列对角占优的 $n\times n$ 矩阵．因而，

$$\sum_{\substack{i=1\\ i\neq j}}^{n}|a_{ij}|<|a_{jj}|\qquad(1\leqslant j\leqslant n)$$

确定不选主元的高斯消元法是否保持它的对角占优性．

41. 在对角占优阵 A 中，用等式

$$e_i=|a_{ii}|-\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|$$

定义**第 i 行的超过量**．说明在定理 5 的证明中下式成立．

$$|a_{ii}-a_{i1}a_{1i}/a_{11}|\geqslant\sum_{\substack{j=2\\ j\neq i}}^{n}|a_{ij}-a_{i1}a_{1j}/a_{11}|+e_i$$

因此，第 i 行的超过量在高斯消元法中不减少．

42. 如果必要的话，参考习题 4.3.32～4.3.34．设 p 是 $\{1,2,\cdots,n\}$ 的一个置换，并设 P 是相应的置换阵．（因此，$P_{ij}=\delta_{p_i j}$．）设 q 是 p 的逆而 Q 是相应于 q 的置换阵．证明 $P^{-1}=Q$.

43.（续）证明：若 P 是置换阵，则 $P^{-1}=P^T$.

44. 若 A 是 $n\times n$ 矩阵，B 是 $n\times m$ 矩阵．用行尺度高斯消元法解 $AX=B$ 需要多少次乘法和除法？当 $B=I$ 时，结果又怎样？

45. 证明或否定：若 A 是三对角阵而 P 是置换阵，则 PAP^{-1} 是三对角阵．

46. 假如在行尺度选主元高斯消元法的每个主步重新计算尺度数组．证明：对一个对称的对角占优阵来说，不选主元的高斯消元法和行尺度选主元高斯消元法是相同的．

47. 证明定理 3.

48. 在行尺度选主元的高斯消元法中，假如尺度数 s_i 由下式定义

$$s_i=|a_{i1}|+|a_{i2}|+\cdots+|a_{in}|$$

证明：如果所得的算法应用于一个对角占优阵时，那么将选取自然的主元次序（1，2，…，n）．编写和测试执行习题 4.3.25 中想法的程序．这是对一个对称方程组不选主元的高斯消元法．

49. 证明或否定猜想：若矩阵有性质

$$0\neq|a_{ii}|\geqslant\sum_{\substack{j=1\\ j\neq i}}|a_{ij}|\qquad(1\leqslant i\leqslant n)$$

则不选主元的高斯消元法将保持这个性质．

185

50. a. 证明用子式展开计算矩阵行列式含有 $(n-1)(n!)$ 次 ops.

b. 证明克拉默法则需要 $(n^2-1)(n!)$ 次 ops.

c. 证明高斯-若尔当方法包含 $\frac{1}{2}n(n+1)^2\approx n^3/2$ 次 ops，所以它比高斯消元法多耗费 50% 的计算量．

计算机习题 4.3

1. 全主元高斯消元法以非自然次序处理行和列．因此，在第 1 步，选主元素 a_{ij} 使得 $|a_{ij}|$ 在整个矩阵中最大．这确定第 i 行是主行而第 j 列是主列．再从其他行中减去第 i 行的倍数在第 j 列中产生 0．编写执行这个过程的算法．这需要两个置换向量．

2. 参考习题 4.3.25，编写一个对对称矩阵执行高斯消元法分解阶段的程序．假定不需要选主元．
3. 编写和测试行尺度选主元高斯消元法的程序．适合的测试情况见习题 4.3.18，4.3.20～4.3.21 和 4.3.25.
4. 编写和测试高斯算法中包含列均衡化的程序(见习题 4.3.5).
5. 编写和测试求解 $Ax=b$ 和 $y^{T}A=c^{T}$ 的程序，只利用 A 的一个分解(带行尺度主元)以及两个求解 x 和 y 的子程序．
6. 编写和测试利用列均衡化，行均衡化和全主元的求解 $Ax=b$ 的程序．(术语参见习题 4.3.5 和上面计算机习题 4.3.1.)
7. 编写用带列均衡化和行尺度的高斯-若尔当方法求解 $n\times n$ 方程组 $Ax=b$ 的子程序 Gaussj(n，A，b，x，p，s，d)，在高斯-若尔当算法(不选主元)第 k 个主步，从所有其他行减去第 k 行的倍数使得 x_k 的系数除了第 k 行外其余所有行都为 0. 最终，矩阵将是对角阵(而不像在高斯消元法中那样是一个上三角阵)．用行尺度选主元，第 p_k 行用作主行，在除了第 p_k 行外所有行中产生 x_k 的 0 系数．正如习题 4.3.6 中讨论过的那样，列均衡化应在开始时就执行．而这个过程中所需要的除数应该存放在数组 d 中，因为最后求 x 时需要它们．
8. 编写和测试尺度高斯消元法的递归形式，其中尺度数组反复重新计算．

4.4 范数和误差分析

讨论数值问题中所涉及的向量误差，通常利用范数．我们的向量通常是$\mathbb{R}^n$空间中的一个向量，但范数可在任何向量空间上定义．

4.4.1 向量范数

在向量空间 V 上，范数是一个从 V 到非负实数集的函数 $\|\cdot\|$，并服从下列三个要求： [186]

$$\text{若 } x\neq 0,\ x\in V,\ \text{则 } \|x\|>0 \tag{1}$$

$$\text{若 } \lambda\in\mathbb{R},\ x\in V,\ \text{则 } \|\lambda x\|=|\lambda|\,\|x\| \tag{2}$$

$$\text{若 } x,\ y\in V,\ \text{则 } \|x+y\|\leqslant\|x\|+\|y\|\ (\text{三角不等式}) \tag{3}$$

我们可以认为 $\|x\|$ 是向量 x 的**长度**或**大小**．向量空间上的范数推广了实数或复数的绝对值概念．$\mathbb{R}^n$上最熟悉的范数是由下式定义的**欧几里得** ℓ_2 **范数**．

$$\|x\|_2=\left(\sum_{i=1}^{n}x_i^2\right)^{1/2},\text{其中 } x=(x_1,x_2,\cdots,x_n)^{T}$$

这个范数对应于我们直觉的长度概念．我们使用下标 2 仅仅作为一个标识符．在数值分析中也使用其他的范数．最简单和最容易计算的范数称为 ℓ_∞ **范数**：

$$\|x\|_\infty=\max_{1\leqslant i\leqslant n}|x_i| \tag{4}$$

此外，使用下标是为了使这个范数区别于其他范数．第 3 种$\mathbb{R}^n$上范数的重要例子称为 ℓ_1 **范数**：

$$\|x\|_1=\sum_{i=1}^{n}|x_i| \tag{5}$$

例 1 利用 $\|\cdot\|_1$ 范数，比较下列$\mathbb{R}^4$中三个向量的长度．然后对 $\|\cdot\|_2$ 范数和 $\|\cdot\|_\infty$ 范数重新计算．

$$x=(4,4,-4,4)^{T}\qquad v=(0,5,5,5)^{T}\qquad w=(6,0,0,0)^{T}$$

解 列出结果如下：

	$\|\cdot\|_1$	$\|\cdot\|_2$	$\|\cdot\|_\infty$
x	16.	8.	4.
v	15.	8.66	5.
w	6.	6.	6.

■

为了更好地理解这些范数，考察$\mathbb{R}^2$是有启发的．对上述三种范数，在图 4-1 中，我们给出集合

$$\{x: x \in \mathbb{R}^2, \|x\| \leqslant 1\}$$

187 的略图．这个集合称为二维向量空间中的**单位单元**或**单位球**.

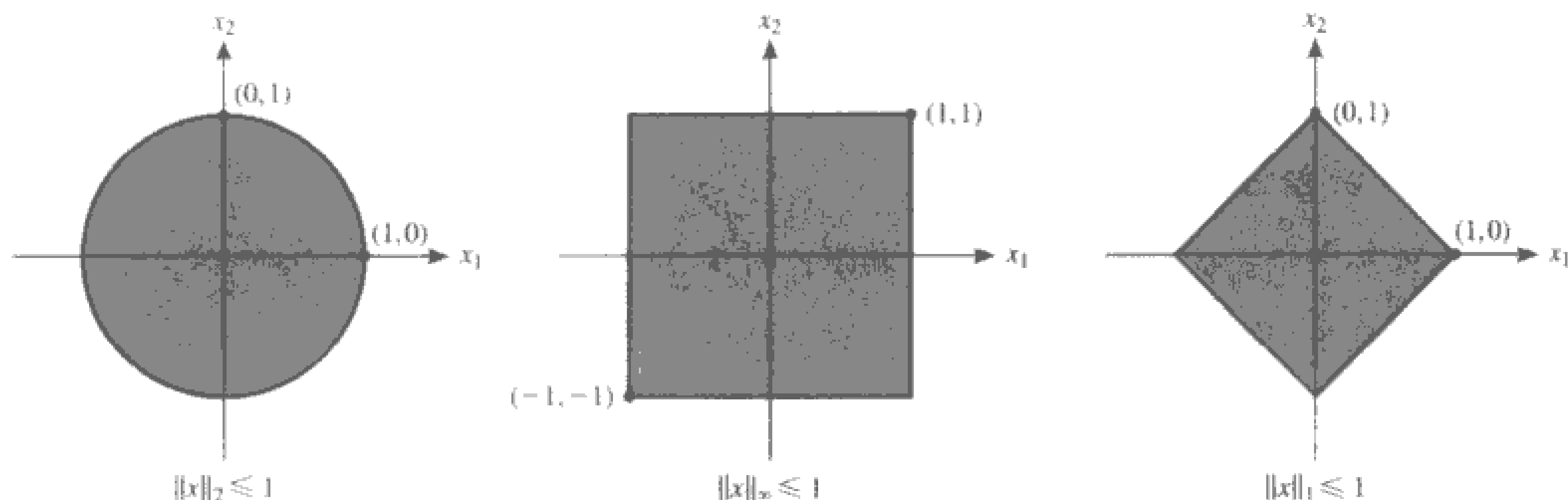


图 4-1 三种范数在$\mathbb{R}^2$ 中的单位单元

4.4.2 矩阵范数

现在我们转到定义矩阵范数的问题．虽然我们可以处理一般的矩阵范数，让它们仅仅服从同样的要求(1)-(3)，但是我们通常更喜欢与一个向量范数密切相关的矩阵范数．若已经指定一个向量范数 $\|\cdot\|$，则**从属于**它的矩阵范数定义为

$$\|A\| = \sup\{\|Au\| : u \in \mathbb{R}^n, \|u\| = 1\} \tag{6}$$

这个范数也称为**对应于**给定向量范数的矩阵范数．这里 A 理解为一个 $n\times n$ 矩阵.

定理 1(从属矩阵范数定理) *若 $\|\cdot\|$ 是$\mathbb{R}^n$上的任意范数，则等式*

$$\|A\| = \sup_{\|u\|=1}\{\|Au\| : u \in \mathbb{R}^n\}$$

定义全体 $n\times n$ 矩阵组成的线性空间上的范数.

证明 我们将验证范数的三条公理．首先，若 $A\neq 0$，则 A 至少有一个非零列，譬如说，$A^{(j)}\neq 0$．考虑第 j 个分量为 1 其余分量为 0 的向量 x；即 $x=(0, \cdots, 0, 1, 0, \cdots, 0)^{\mathrm{T}}$．显然，$x\neq 0$ 且向量 $v=x/\|x\|$ 的范数为 1．因此，由 $\|A\|$ 的定义

$$\|A\| \geqslant \|Av\| = \frac{\|Ax\|}{\|x\|} = \frac{\|A^{(j)}\|}{\|x\|} > 0$$

其次，从向量范数的性质(2)，我们有

$$\|\lambda A\| = \sup\{\|\lambda Au\| : \|u\| = 1\} = |\lambda| \sup\{\|Au\| : \|u\| = 1\} = |\lambda| \|A\|$$

188

关于三角不等式，我们利用向量范数类似的性质和习题 4.4.4 可得

$$\begin{aligned}\|A+B\| &= \sup\{\|(A+B)u\| : \|u\| = 1\}\\ &\leqslant \sup\{\|Au\| + \|Bu\| : \|u\| = 1\}\\ &\leqslant \sup\{\|Au\| : \|u\| = 1\} + \sup\{\|Bu\| : \|u\| = 1\}\\ &= \|A\| + \|B\|\end{aligned}$$

■

定义(6)的一个重要推论，并且确实由定义导出的是

$$\|Ax\| \leqslant \|A\| \|x\| \qquad (x \in \mathbb{R}^n) \tag{7}$$

为证明这个结论，我们看到对 $x=0$，它显然成立. 若 $x\neq 0$，则向量 $v=x/\|x\|$ 的范数为 1. 因此，由(6)式可得

$$\|A\| \geqslant \|Av\| = \frac{\|Ax\|}{\|x\|}$$

作为这个重要概念的一个实例，取定我们的向量范数为(4)式所定义的 $\|x\|_\infty$ 范数. 试问它的从属矩阵范数是什么？下面是计算：

$$\begin{aligned}\|A\|_\infty &= \sup_{\|u\|_\infty=1} \|Au\|_\infty\\ &= \sup_{\|u\|_\infty=1}\left\{\max_{1\leqslant i\leqslant n} |(Au)_i|\right\} = \max_{1\leqslant i\leqslant n}\left\{\sup_{\|u\|_\infty=1} |(Au)_i|\right\}\\ &= \max_{1\leqslant i\leqslant n}\left\{\sup_{\|u\|_\infty=1}\left|\sum_{j=1}^n a_{ij}u_j\right|\right\} = \max_{1\leqslant i\leqslant n}\sum_{j=1}^n |a_{ij}|\end{aligned} \tag{8}$$

这里我们利用了两个最大化过程可以交换(习题 4.4.9)，也利用了下列事实：对固定的 i 和 $\|u\|_\infty=1$，当 $a_{ij}\geqslant 0$ 取 $u_j=+1$ 以及 $a_{ij}<0$ 时取 $u_j=-1$ 可以得到 $\left|\sum_{j=1}^n a_{ij}u_j\right|$ 的上确界.

因此，我们已证得下列定理.

定理 2(无穷矩阵范数定理) 若向量范数 $\|\cdot\|_\infty$ 由下式定义：

$$\|x\|_\infty = \max_{1\leqslant i\leqslant n} |x_i|$$

则它的从属矩阵范数由下式给出：

$$\|A\|_\infty = \max_{1\leqslant i\leqslant n}\sum_{j=1}^n |a_{ij}|$$

189

从属于一个向量范数的矩阵范数除了具有基本性质(1)～(3)外还有其他的性质. 例如，

$$\|I\| = 1 \tag{9}$$

$$\|AB\| \leqslant \|A\| \|B\| \tag{10}$$

(9)式的证明可直接从(6)式可得，而不等式(10)从(6)式和不等式(7)可得.

另一个重要的矩阵范数是由下式定义的 ℓ_2 **矩阵范数**(也称**谱范数**)：

$$\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2$$

这是从属于欧几里得向量范数的矩阵范数. 在 5.4 节的定理 5 中，我们将证明

$$\| A \|_2 = \max_{1 \leqslant i \leqslant n} | \sigma_i |$$

其中 σ_i 是 A 的**奇异值**. 若 $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_n \geqslant 0$，则从 5.4 节定理 1 的证明我们有 $Av_1 = \sigma_1 u_1$ 和 $A^T u_1 = \sigma_1 v_1$. 因此 $A^T A v_1 = \sigma_1^2 v_1$，故 σ_1^2 等于 $A^T A$ 的最大特征值. 因此，2 矩阵范数经常被定义为

$$\| A \|_2 = \sqrt{\rho(A^T A)}$$

其中 $\rho(A^T A)$ 称为 $A^T A$ 的**谱半径**，且被定义为 $A^T A$ 的最大特征值.

4.4.3 条件数

下面就把这些概念用于计算. 我们考虑方程

$$Ax = b$$

其中 A 是 $n \times n$ 矩阵. 假定 A 可逆.

例 2 若扰动 A^{-1} 后得到一个新矩阵 B，则扰动解 $x = A^{-1}b$ 后变成一个新向量 $\tilde{x} = Bb$. 试问后面扰动的绝对和相对界有多大?

解 利用向量范数和它的从属矩阵范数，我们有

$$\| x - \tilde{x} \| = \| x - Bb \| = \| x - BAx \| = \| (I - BA)x \| \leqslant \| I - BA \| \, \| x \|$$

[190] 这给出 x 扰动的大小. 若度量相对扰动，我们可以写成

$$\frac{\| x - \tilde{x} \|}{\| x \|} \leqslant \| I - BA \| \tag{11}$$

不等式(11)给出 $\| x - \tilde{x} \| / \| x \|$ 的一个上界，而取这个比为 x 和 $\tilde{x}$ 之间**相对误差**的一个度量. ■

例 3 假如向量 b 扰动后得到一个向量 $\tilde{b}$. 若 x 和 $\tilde{x}$ 满足 $Ax = b$ 和 $A\tilde{x} = \tilde{b}$. 试问 x 和 $\tilde{x}$ 差的绝对和相对界有多大?

解 假设 A 可逆，我们有

$$\| x - \tilde{x} \| = \| A^{-1}b - A^{-1}\tilde{b} \| = \| A^{-1}(b - \tilde{b}) \| \leqslant \| A^{-1} \| \, \| b - \tilde{b} \|$$

这给出了 x 的扰动范围. 为了估计相对扰动，我们写

$$\| x - \tilde{x} \| \leqslant \| A^{-1} \| \, \| b - \tilde{b} \| = \| A^{-1} \| \, \| Ax \| \frac{\| b - \tilde{b} \|}{\| b \|} \leqslant \| A^{-1} \| \, \| A \| \, \| x \| \frac{\| b - \tilde{b} \|}{\| b \|}$$

因此，

$$\frac{\| x - \tilde{x} \|}{\| x \|} \leqslant \kappa(A) \frac{\| b - \tilde{b} \|}{\| b \|} \tag{12}$$

其中

$$\kappa(A) \equiv \| A \| \cdot \| A^{-1} \| \tag{13}$$

数 $\kappa(A)$ 称为矩阵 A 的**条件数**. ■

不等式(12)告诉我们 x 的相对误差不大于 $\kappa(A)$ 乘 b 的相对误差. 条件数与开始分析时所选取的向量范数有关. 从不等式(12)，我们看到当条件数小时，b 的小扰动只导致 x 的小扰动. 而不等式 $\kappa(A) \geqslant 1$ 总成立. (见习题 4.4.14)

下面是一个说明条件数的例子：设 $\varepsilon > 0$ 且

$$A = \begin{bmatrix} 1 & 1+\varepsilon \\ 1-\varepsilon & 1 \end{bmatrix} \qquad A^{-1} = \varepsilon^{-2} \begin{bmatrix} 1 & -1-\varepsilon \\ -1+\varepsilon & 1 \end{bmatrix}$$

若使用∞范数，则由(8)式我们有$\| A \|_\infty = 2+\varepsilon$，$\| A^{-1} \|_\infty = \varepsilon^{-2}(2+\varepsilon)$．因此，$\kappa(A) = [(2+\varepsilon)/\varepsilon]^2 > 4/\varepsilon^2$．若$\varepsilon \leqslant 0.01$，则$\kappa(A) \geqslant 40\ 000$．此时，相对$b$的一个小扰动可能就会导致方程组$Ax=b$的解有一个超过40 000倍的相对扰动．

191

如果我们求解方程组

$$Ax = b$$

数值上讲，我们未得到精确解x而得到近似解$\tilde{x}$．我们可形成$A\tilde{x}$来检验$\tilde{x}$，看看$A\tilde{x}$是否接近于b．从而，我们得到**残差向量**

$$r = b - A\tilde{x}$$

精确解x和近似解$\tilde{x}$之差称为**误差向量**

$$e = x - \tilde{x}$$

下列误差向量和残差向量之间的关系

$$Ae = r \tag{14}$$

是十分重要的．

注意$\tilde{x}$是线性方程组$A\tilde{x}=\tilde{b}$的精确解，它有一个扰动的右端项$\tilde{b}=b-r$．我们现在建立x与b相对误差间的关系．换言之，我们要给出$\| x-\tilde{x} \| / \| x \|$和比率$\| b-\tilde{b} \| / \| b \| = \| r \| / \| b \|$之间的关系．下列定理表明条件数$\kappa(A)$在其中扮演了一个重要的角色．

定理 3（包含条件数界的定理） *在解方程组$Ax=b$中，条件数$\kappa(A)$，残差向量r和误差向量e满足下列不等式：*

$$\frac{1}{\kappa(A)} \frac{\| r \|}{\| b \|} \leqslant \frac{\| e \|}{\| x \|} \leqslant \kappa(A) \frac{\| r \|}{\| b \|}$$

证明 右边的不等式可写成

$$\| e \| \| b \| \leqslant \| A \| \| A^{-1} \| \| r \| \| x \|$$

这是成立的，因为

$$\| e \| \| b \| = \| A^{-1} r \| \| Ax \| \leqslant \| A^{-1} \| \| r \| \| A \| \| x \|$$

（事实上，本定理中右边的不等式就是不等式(12)）．左边的不等式可写成

$$\| r \| \| x \| \leqslant \| A \| \| A^{-1} \| \| b \| \| e \|$$

而此式又可从下式立即得到

$$\| r \| \| x \| = \| Ae \| \| A^{-1} b \| \leqslant \| A \| \| e \| \| A^{-1} \| \| b \|$$

■ 192

一个具有大条件数的矩阵称为**病态的**．而一个病态矩阵A会出现方程组$Ax=b$的解对向量b的小变化非常敏感的情况．换言之，为使所求的x达到某种精度，我们应该要求b有足够高的精度．若A的条件数大小适中的话，就称该矩阵为**良态的**．

习题 4.4

1. 证明范数$\| x \|_\infty$，$\| x \|_2$和$\| x \|_1$满足范数的条件(1)，(2)，(3)．
2. 证明$\| x \|_\infty \leqslant \| x \|_2 \leqslant \| x \|_1$，对一切$x \in \mathbb{R}^n$，并且说明即使对非零向量等式也有可能成立．
3. （续）证明$\| x \|_1 \leqslant n \| x \|_\infty$和$\| x \|_2 \leqslant \sqrt{n} \| x \|_\infty$，$x \in \mathbb{R}^n$．
4. 说明对任何两个到$\mathbb{R}$内的函数f和g，有

$$\sup[f(x)+g(x)] \leqslant \sup f(x) + \sup g(x)$$

5. 对定理 2 中的矩阵范数，证明或否定：$\|AB\|_\infty = \|A\|_\infty \|B\|_\infty$. 特殊情况 $\|A^2\|_\infty = \|A\|_\infty^2$ 怎么样？
6. 说明在$\mathbb{R}^n$上定义的范数必须以某种方式包括一个向量的所有分量. 说明范数 $\|x\|_\infty$ 确实是$\mathbb{R}^n$上最简单的范数.（你必须定义合适的简单性概念.）
7. 确定下列表达式是否定义在$\mathbb{R}^n$上的范数：

   a. $\max\{|x_2|, |x_3|, \cdots, |x_n|\}$

   b. $\sum_{i=1}^{n} |x_i|^3$

   c. $\left\{\sum_{i=1}^{n} |x_i|^{1/2}\right\}^2$

   d. $\max\{|x_1-x_2|, |x_1+x_2|, |x_3|, |x_4|, \cdots, |x_n|\}$

   e. $\sum_{i=1}^{n} 2^{-i} |x_i|$
8. 定义

$$\|A\| = \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|$$

   说明这是一个范数(即在全体 $n\times n$ 矩阵组成的线性空间上的范数). 并说明它不从属于任何向量范数. 试问它符合(9)式及不等式(10)吗？
9. a. 证明：若 A 和 B 是任意集合，f 是 $A\times B$ 上的一个有界实函数，则

$$\sup_{a\in A}\sup_{b\in B} f(a,b) = \sup_{b\in B}\sup_{a\in A} f(a,b)$$

   b. 举例说明，通常一个上确界和一个下确界是不能交换的.

   c. 说明

$$\sup_{a\in A}\inf_{b\in B} f(a,b) \leqslant \inf_{b\in B}\sup_{a\in A} f(a,b)$$

10. 证明：对任何向量范数及其从属矩阵范数，以及对任何 $n\times n$ 矩阵 A，存在相应的一个向量 $x\neq 0$ 使得 $\|Ax\| = \|A\|\,\|x\|$.

193

11. 说明对(5)式定义的向量范数 $\|x\|_1$，从属的矩阵范数是

$$\|A\|_1 = \max_{1\leqslant j\leqslant n}\sum_{i=1}^{n} |a_{ij}|$$

12. (续)利用上题中的矩阵范数，计算矩阵$\begin{bmatrix} 1 & 1+\varepsilon \\ 1-\varepsilon & 1 \end{bmatrix}$的条件数.
13. 请问从属矩阵范数满足 $\|AB\| = \|BA\|$ 吗？说明理由.
14. 证明可逆阵的条件数至少为 1.
15. 试问什么样的矩阵其条件数等于 1？
16. 利用(8)式中的无穷矩阵范数，计算矩阵$\begin{bmatrix} 7 & 8 \\ 9 & 10 \end{bmatrix}$的条件数.
17. 设 A 是 $n\times n$ 矩阵，其逆为 $C=(c_{ij})$. 说明在 $Ax=b$ 的解中，b_j 的扰动量 δ 将引起 x_i 的扰动 $c_{ij}\delta$.
18. (续)证明：a_{jk} 的扰动量 δ 将产生 x_i 的近似扰动 $-c_{ij}x_k\delta$.
19. (续)下列等式有时被用作为一个 $n\times n$ 矩阵 A 的条件数：

$$M(A) = n \max_{1\leqslant i,j\leqslant n} |a_{ij}| \max_{1\leqslant i,j\leqslant n} |c_{ij}|$$

   其中 C 是 A 的逆. 证明：若 $Ax=b$ 且只扰动了 b 的一个分量(譬如说数量为 ε)，则扰动后的 x（记为 $\tilde{x}$）满足

$$\frac{\| x-\widetilde{x} \|_\infty}{\| x \|_\infty} \leqslant M(A) \frac{|\epsilon|}{\| b \|_\infty}$$

20. 在课本中已证明：若 $Ax=b$ 和 $A\widetilde{x}=\tilde{b}$，则

$$\frac{\| x-\widetilde{x} \|}{\| x \|} \leqslant \kappa(A) \frac{\| b-\tilde{b} \|}{\| b \|}$$

证明：对每个非奇异阵 A，这个不等式对某些向量 b 和 $\tilde{b}$ 会变成一个等式.（当然，我们要求 $b\neq 0$ 且 $b\neq \tilde{b}$.）提示：完成上述不等式的证明并找出等式出现的条件.

21. 设 $n=3$ 并设

$$A=\begin{bmatrix} 4 & -3 & 2 \\ -1 & 0 & 5 \\ 2 & 6 & -2 \end{bmatrix}$$

在所有满足 $\| x \|_\infty \leqslant 1$ 的向量 x 中，求出一个向量使得 $\| Ax \|_\infty$ 尽可能地大. 再给出 $\| A \|_\infty$ 的数值.

22. $\mathbb{R}^n$ 上的**加权 ℓ_∞ 范数**是一个形如

$$\| x \| = \max_{1\leqslant i\leqslant n} w_i \mid x_i \mid$$

的范数，这里 w_1，w_2，…，w_n 是固定的正数称为权. 证明这个范数满足范数基本条件. 那么其从属的矩阵范数是什么？

23. 证明：若 $\| \cdot \|$ 是向量空间上的一个范数，并且如果我们用某个固定的正数 α 定义 $\| x \|'=\alpha \| x \|$，则 $\| \cdot \|'$ 也是一个范数.

194

24. （续）若将上题中的说法应用于从属矩阵范数，则所得的范数也是一个从属矩阵范数.

25. 设 $\| \cdot \|$ 是向量空间 V 上的一个范数，对 V 中的 x 和 y，设 $d(x, y)=\| x-y \|$，说明 d 具有下列性质：
    a. $d(x, x)=0$
    b. $d(x, y)=d(y, x)$
    c. 若 $x\neq y$，则 $d(x, y)>0$
    d. $d(x, y)\leqslant d(x, z)+d(z, y)$

    （具有这四个性质的函数称为**度规**.）

26. 举出一个行列式非常小的良态矩阵的例子.

27. 计算对角线上为 $+1$，对角线下面为 -1 的 $n\times n$ 下三角阵的条件数. 采用矩阵范数 $\| \cdot \|_\infty$.

28. 证明：若 A 有非平凡的不动点(即 $Ax=x\neq 0$)，则对任何从属矩阵范数 $\| A \| \geqslant 1$.

29. 在 $\mathbb{R}^2$ 上举一个范数例子使(1，0)的范数是 2 而(1，1)的范数是 1.

30. 在 $\mathbb{R}^2$ 中是否存在一个使 $\| (1,0) \| = \| (0,1) \| = \| (1/3, 1/3) \|$ 的范数？

31. （见习题 4.4.2～4.4.3）求 x 的精确条件使得
    a. $\| x \|_\infty = \| x \|_1$
    b. $\| x \|_\infty = \| x \|_2$
    c. $\| x \|_1 = \| x \|_2$

32. 说明 $\| A \|$ 是对一切 x 使得 $\| Ax \| \leqslant M \| x \|$ 成立的最小的数 M.

33. 对任意 $n\times n$ 矩阵 A，定义

$$\| A \|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2\right)^{1/2}$$

（这个范数称为**弗罗贝尼乌斯范数**.）这个范数是否为从属矩阵范数？对 $\| A \| = \max\limits_{1\leqslant i,j\leqslant n} \mid a_{ij} \mid$ 回答同样的问题. 证明这些等式定义了全体 $n\times n$ 矩阵组成的向量空间上的范数.

34. 对任意从属矩阵范数，置换阵的范数为 1 必定成立吗？说明理由.

35. 对向量 $x=(x_1, x_2, \cdots, x_n)^T \in \mathbb{R}^n$，我们定义向量 x 的绝对值 $|x|$ 为向量 $(|x_1|, |x_2|, \cdots, |x_n|)^T$. 对向量 x 和 y，我们也定义 $x \leqslant y$，意指对 $i=1, 2, \cdots, n$，$x_i \leqslant y_i$. 证明范数 $\|\cdot\|_1$，$\|\cdot\|_2$，$\|\cdot\|_\infty$ 具有下列性质：若 $|x| \leqslant |y|$，则 $\|x\| \leqslant \|y\|$.

36. 对任何实数 $p \geqslant 1$，公式

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$$

定义一个范数(证明查阅 Bartle[1976，第 60 页].) 证明：对每个 $x \in \mathbb{R}^n$，

$$\lim_{p \to \infty} \|x\|_p = \|x\|_\infty$$

195 这说明了为什么要使用记号 $\|\cdot\|_\infty$.

37. 证明下列范数的性质：

a. $\|0\|=0$

b. $\|x+y\| \geqslant |\|x\|-\|y\||$

c. 对向量 $x^{(1)}, x^{(2)}, \cdots, x^{(m)}$，$\left\|\sum_{i=1}^{m} x^{(i)}\right\| \leqslant \sum_{i=1}^{m} \|x^{(i)}\|$

38. 设 $\|\cdot\|$ 是 $\mathbb{R}^n$ 上的一个范数. 定义

$$\|x\|' = \sup\{u^T x : u \in \mathbb{R}^n, \|u\| = 1\}$$

证明这个等式定义一个范数. 证明：若重复这个过程，则得到原来的范数；即 $(\|\cdot\|')' = \|\cdot\|$. 证明对一切 $x, y \in \mathbb{R}^n$，

$$|x^T y| \leqslant \|x\| \|y\|'$$

39. 证明(13)式中定义的关于条件数的不等式

$$\kappa(AB) \leqslant \kappa(A)\kappa(B)$$

40. 利用 $\|A\|_1$，$\|A\|_2$，$\|A\|_\infty$ 计算条件数

a. $\begin{bmatrix} a+1 & a \\ a & a-1 \end{bmatrix}$

b. $\begin{bmatrix} 0 & 1 \\ -2 & 0 \end{bmatrix}$

c. $\begin{bmatrix} \alpha & 1 \\ 1 & 1 \end{bmatrix}$

41. 利用习题 4.4.2～4.4.3，证明

$$n^{-1}\|A\|_2 \leqslant n^{-1/2}\|A\|_\infty \leqslant \|A\|_2 \leqslant n^{1/2}\|A\|_1 \leqslant n\|A\|_2$$

42. 对矩阵 $A=\begin{bmatrix} 1 & 2 \\ 1 & 2.01 \end{bmatrix}$ 求解方程组 $Ax=b$. 估计 b 中多小的改变将会影响到解 x. 当 $b=(4, 4)$ 和 $b'=(3, 5)$ 时检验你的预测.

43. 设 A 是 $m \times n$ 矩阵. 我们理解 A 为从具有范数 $\|\cdot\|_1$ 的 $\mathbb{R}^n$ 到具有范数 $\|\cdot\|_\infty$ 的 $\mathbb{R}^m$ 的一个线性映射. 在这些情况下 $\|A\|$ 是什么？下列简单公式

$$\|A\| = \max\{\|Ax\|_\infty : \|x\|_1 = 1\}$$

就是我们所想要的.

44. (续)当范数 $\|\cdot\|_1$ 和 $\|\cdot\|_\infty$ 交换后，试着讨论上题.

45. 证明条件数 $\kappa(A)$ 可用下列公式表示

$$\kappa(A)=\sup_{\|x\|=\|y\|}\|Ax\|/\|Ay\|$$

46. 设 $\|\cdot\|$ 是 $\mathbb{R}^n$ 上的范数，A 是 $n\times n$ 矩阵．取 $\|x\|'=\|Ax\|$．为保证 $\|\cdot\|'$ 也是一个范数，试问 A 的明确条件是什么？
47. 证明如果使用欧几里得范数，那么集合

$$H=\{x\in\mathbb{R}^n:\|x-a\|=\|x-b\|\}$$

就是一个超平面(即解释为一个 $n-1$ 维的线性子空间)，但是这个结论对其他范数一般不成立．以 $n=2$ 的情况举例说明．

196

48. 证明条件数有性质

$$\kappa(\lambda A)=\kappa(A)\qquad(\lambda\neq 0)$$

49. 证明：若方阵 A 对一切 x 及 $\theta>0$ 满足不等式 $\|Ax\|\geqslant\theta\|x\|$，则 A 非奇异且 $\|A^{-1}\|\leqslant\theta^{-1}$．这个结论对任何向量范数及其从属矩阵范数都成立．
50. (续)证明：若 A 对角占优，则它具有上题中的性质．当使用 $\|\cdot\|_\infty$ 范数时给出 θ 的一个值．
51. 证明：若 A 非奇异，则存在一个 $\delta>0$ 使得对一切满足 $\|E\|<\delta$ 的矩阵 E，矩阵 $A+E$ 非奇异．可以证明对全体 $n\times n$ 矩阵组成的向量空间上的任何范数这个结论也成立．
52. (续)说明在上题中我们可使用

$$\delta=\inf\{\|Ax\|:\|x\|=1\}$$

这里使用了一个向量范数及其从属矩阵范数．

53. 设 A 是 $n\times n$ 矩阵而 N 是它的零空间(核)．定义

$$\delta=\inf\{\|Ax\|:\|x\|=1\text{ 且 }x\perp N\}$$

证明：若 $\|E\|<\delta$，则 $\operatorname{rank}(A+E)\geqslant\operatorname{rank}(A)$．

54. 证明：若 A 非奇异，则存在一个奇异阵 B 使得 $\|B-A\|_2=\|A^{-1}\|_2$．
55. 证明(14)式．
56. 证明

$$\|A\|_2=\max_{\substack{\|x\|_2=1\\ \|y\|_2=1}}|y^TAx|$$

计算机习题 4.4

编写计算向量和方阵范数的子程序．使用**极大范数** $\|x\|_\infty$ 及其从属矩阵范数．

4.5 诺伊曼级数和迭代细化

范数的一个重要应用出现在精确的处理向量空间中的收敛性概念．若向量空间 V 指定一个范数 $\|\cdot\|$，则一对(V，$\|\cdot\|$)就是一个**赋范线性空间**．向量序列 $v^{(1)}$，$v^{(2)}$，…收敛的概念定义如下：如果

$$\lim_{k\to\infty}\|v^{(k)}-v\|=0$$

那么我们说给定的序列**收敛**于向量 v．这和我们的直觉想法一致，即当 k 增加时，向量 $v^{(k)}$ 和极限向量 v 之间的距离接近于 0．

197

下面是一个在$\mathbb{R}^4$ 中的例子：设

$$v^{(k)} = \begin{bmatrix} 3 - k^{-1} \\ -2 + k^{-\frac{1}{2}} \\ (k+1)k^{-1} \\ e^{-k} \end{bmatrix} \text{和} v = \begin{bmatrix} 3 \\ -2 \\ 1 \\ 0 \end{bmatrix}$$

则

$$v^{(k)} - v = \begin{bmatrix} -k^{-1} \\ k^{-\frac{1}{2}} \\ k^{-1} \\ e^{-k} \end{bmatrix}$$

如果我们利用 4.4 节的无穷范数计算$\| v^{(k)} - v \|$，那么我们就会看到当$k \to \infty$时，$\| v^{(k)} - v \|_\infty \to 0$. 因此，在赋范线性空间$(\mathbb{R}^4, \| \cdot \|_\infty)$中$v$是序列$[v^{(k)}]$的极限.

现在不加证明地给出一个关于赋范线性空间的重要结论是合适的：在有限维向量空间上任何两个范数都会导致相同的收敛性概念. 因此，在上例中，已验证$\| v^{(k)} - v \|_\infty \to 0$，我们不用再计算就能得到对$\mathbb{R}^4$中任何的范数有$\| v^{(k)} - v \| \to 0$. 警告：这个定理在无限维赋范线性空间上不能应用.（例如，见习题 4.5.20.）

另一个关于有限维赋范线性空间的重要结果是在这样的一个空间中，每个柯西序列收敛. 因此，若一个有限维赋范线性空间中的序列$[v^{(k)}]$满足**柯西准则**

$$\lim_{k \to \infty} \sup_{i,j \geqslant k} \| v^{(i)} - v^{(j)} \| = 0$$

则必定存在一个点$v \in V$使得序列收敛于它.

我们将对$\mathbb{R}^n$中的向量和$n \times n$矩阵应用这些概念. 在下面的定理中，我们取$\| \cdot \|$为$\mathbb{R}^n$中的任意范数并利用 4.4 节中定义的其从属矩阵范数.

定理 1（诺伊曼级数定理） 若A是使$\| A \| < 1$的$n \times n$矩阵，则$I - A$可逆且

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k \tag{1}$$

证明 首先，我们将证明$I - A$可逆. 若它不可逆，则它奇异，且存在一个向量x满足$\| x \| = 1$和$(I - A)x = 0$. 由此我们有

$$1 = \| x \| = \| Ax \| \leqslant \| A \| \, \| x \| = \| A \|$$

[198] 这就与假设$\| A \| < 1$矛盾了. 下面，我们将证明诺伊曼级数的部分和收敛于$(I - A)^{-1}$：

$$\sum_{k=0}^{m} A^k \to (I - A)^{-1} \qquad (\text{当 } m \to \infty)$$

为此只需证明

$$(I - A) \sum_{k=0}^{m} A^k \to I \qquad (\text{当 } m \to \infty) \tag{2}$$

即可. 左边可写成

$$(I - A) \sum_{k=0}^{m} A^k = \sum_{k=0}^{m} (A^k - A^{k+1}) = A^0 - A^{m+1} = I - A^{m+1}$$

因为当 $m\to\infty$ 时，$\|A^{m+1}\|\leqslant\|A\|^{m+1}\to 0$，所以这就证明了(2)式.（为什么?） ■

在任何巴拿赫空间上的连续线性算子理论中这个定理本质上以同样的形式出现. 本定理在实际应用和理论推导两方面都非常重要. 观察(1)式，我们得到估计：

$$\|(I-A)^{-1}\|\leqslant\sum_{k=0}^{\infty}\|A^k\|\leqslant\sum_{k=0}^{\infty}\|A\|^k=\frac{1}{1-\|A\|}$$

例 1 利用诺伊曼级数计算下列矩阵之逆.

$$B=\begin{bmatrix}0.9 & -0.2 & -0.3\\ 0.1 & 1.0 & -0.1\\ 0.3 & 0.2 & 1.1\end{bmatrix}$$

解 设 $B=I-A$，其中

$$A=\begin{bmatrix}0.1 & 0.2 & 0.3\\ -0.1 & 0.0 & 0.1\\ -0.3 & -0.2 & -0.1\end{bmatrix}$$

因为 $\|A\|_\infty=0.6$，所以诺伊曼级数 $\sum_{k=0}^{\infty}A^k$ 收敛于 B^{-1}. 利用习题 4.5.22 中的算法，我们计算某些部分和：

$$\sum_{k=0}^{0}A^k=\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$$

$$\sum_{k=0}^{1}A^k=\begin{bmatrix}1.1 & 0.2 & 0.3\\ -0.1 & 1.0 & 0.1\\ -0.3 & -0.2 & 0.9\end{bmatrix}$$

199

$$\sum_{k=0}^{2}A^k=\begin{bmatrix}1.00 & 0.16 & 0.32\\ -0.14 & 0.96 & 0.06\\ -0.28 & -0.24 & 0.80\end{bmatrix}$$

$$\vdots$$

$$\sum_{k=0}^{19}A^k=\begin{bmatrix}1.000\,000\,00 & 0.142\,857\,14 & 0.285\,714\,29\\ -0.125\,000\,00 & 0.964\,285\,71 & 0.053\,571\,43\\ -0.250\,000\,00 & -0.214\,285\,71 & 0.821\,428\,57\end{bmatrix}$$

最后的部分和给出精确到 8 位小数的 B^{-1}. ■

下面是诺伊曼级数定理的一个变形.

定理 2(可逆阵定理) *若 A 和 B 是使 $\|I-AB\|<1$ 的两个 $n\times n$ 矩阵，则 A 和 B 可逆. 进而，我们有*

$$A^{-1}=B\sum_{k=0}^{\infty}(I-AB)^k \quad 和 \quad B^{-1}=\sum_{k=0}^{\infty}(I-AB)^kA \tag{3}$$

证明 由前面的定理得，AB 可逆且其逆为

$$(AB)^{-1}=\sum_{k=0}^{\infty}(1-AB)^k$$

因此，

$$A^{-1} = BB^{-1}A^{-1} = B(AB)^{-1} = B\sum_{k=0}^{\infty}(I-AB)^k$$

$$B^{-1} = B^{-1}A^{-1}A = (AB)^{-1}A = \sum_{k=0}^{\infty}(I-AB)^k A$$ ■

4.5.1 迭代细化

若 $x^{(0)}$ 是方程

$$Ax = b$$

的近似解，则精确解 x 为

$$x = x^{(0)} + A^{-1}(b - Ax^{(0)}) = x^{(0)} + e^{(0)} \tag{4}$$

[200] 其中 $e^{(0)} = A^{-1}(b - Ax^{(0)})$ 称为**误差向量**. 对应于近似解 $x^{(0)}$ 的**残差向量**是 $r^{(0)} = b - Ax^{(0)}$. 它是可计算的. 当然，我们不计算 A^{-1}，但是向量 $e^{(0)} = A^{-1}r^{(0)}$ 可以通过解方程

$$Ae^{(0)} = r^{(0)}$$

得到. 这些注释导致一个称为**迭代改进**或**迭代细化**的数值过程，下面我们更详细地描述它.

假如方程 $Ax=b$ 已用 4.3 节的高斯消元法解出. 由于舍入误差，不能期望所得结果是精确解，我们用 $x^{(0)}$ 来表示它，然后计算 $r^{(0)}$，$e^{(0)}$ 并用下面三个式子计算 $x^{(1)}$，

$$\begin{cases} r^{(0)} = b - Ax^{(0)} \\ Ae^{(0)} = r^{(0)} \\ x^{(1)} = x^{(0)} + e^{(0)} \end{cases} \tag{5}$$

为得到较好的解 $x^{(2)}$，$x^{(3)}$，…，这个过程可以重复. 该方法的成功依赖于用双精度计算残差 $r^{(i)}$ 以避免在减法中丢失所期望的有效数字. 因此，表达式 $b_i - \sum_{j=1}^{n} a_{ij}x_j^{(0)}$ 用双精度求值.（记住，理论上，$r^{(i)}$ 应该是零向量，所以计算 $r^{(i)}$ 中所涉及的减法必定包含几乎相等的量.）

例 2 应用迭代细化解下列方程组

$$\begin{bmatrix} 420 & 210 & 140 & 105 \\ 210 & 140 & 105 & 84 \\ 140 & 105 & 84 & 70 \\ 105 & 84 & 70 & 60 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 875 \\ 539 \\ 399 \\ 319 \end{bmatrix}$$

解 首先，用行尺度主元高斯消元法分解方程组，并从向后回代得到解

$$x^{(0)} = (0.999\,988, 1.000\,137, 0.999\,670, 1.000\,215)^T$$

由迭代改进的若干步给出

$$x^{(1)} = (0.999\,994, 1.000\,069, 0.999\,831, 1.000\,110)^T$$

$$x^{(2)} = (0.999\,996, 1.000\,046, 0.999\,891, 1.000\,070)^T$$

$$x^{(3)} = (0.999\,993, 1.000\,080, 0.999\,812, 1.000\,121)^T$$

$$x^{(4)} = (1.000\,000, 1.000\,006, 0.999\,984, 1.000\,011)^T$$

真解是 $x=(1,1,1,1)^T$. 因为执行这些计算的计算机近似地具有 Marc-32 的精度，所以我们认为这最后的结果十分良好. ■ 201

为从理论上分析这个算法，我们采用下列观点，即我们的解 $x^{(0)}$ 是由公式

$$x^{(0)} = Bb$$

得到的，其中 B 是 A 的一个近似逆. 于是迭代过程可写成

$$x^{(k+1)} = x^{(k)} + B(b - Ax^{(k)}) \qquad (k \geqslant 0) \tag{6}$$

我们将指出这些向量是诺伊曼级数的部分和，并且利用这个事实去说明序列[$\{x^{(k)}$]收敛于 $Ax=b$ 的解.

我们要说明不精确的措词"B 是 A 的一个近似逆"意指 $\|I-AB\|<1$. 由定理 2，A^{-1}可由下式给出

$$A^{-1} = B\sum_{k=0}^{\infty}(I-AB)^k \tag{7}$$

因此，方程 $Ax=b$ 的精确解是

$$x = B\sum_{k=0}^{\infty}(I-AB)^k b \tag{8}$$

定理 3(迭代改进定理) *若 $\|I-AB\|<1$，则由(6)式给出的迭代改进方法产生向量序列*

$$x^{(m)} = B\sum_{k=0}^{m}(I-AB)^k b \qquad (m \geqslant 0)$$

这些是(8)式中的部分和，所以收敛于 x.

证明 我们使用归纳法. 因为 $x^{(0)}=Bb$，所以 $m=0$ 的情况成立. 若假定第 m 次情况成立，则对第$(m+1)$次情况也成立，因为

$$\begin{aligned} x^{(m+1)} &= x^{(m)} + B(b - Ax^{(m)}) \\ &= B\sum_{k=0}^{m}(I-AB)^k b + Bb - BAB\sum_{k=0}^{m}(I-AB)^k b \\ &= B\left\{b + (I-AB)\sum_{k=0}^{m}(I-AB)^k b\right\} = B\sum_{k=0}^{m+1}(I-AB)^k b \end{aligned}$$

■

也可以直接证明向量 $x^{(m)}$收敛于 x. 为此，利用(6)式得到

$$\begin{aligned} x^{(m+1)} - x &= x^{(m)} - x + B(Ax - Ax^{(m)}) \\ &= (I-BA)(x^{(m)} - x) \end{aligned}$$

202

由此可得

$$\begin{aligned} \|x^{(m+1)} - x\| &\leqslant \|I-BA\| \, \|x^{(m)} - x\| \\ &\leqslant \|I-BA\|^2 \, \|x^{(m-1)} - x\| \\ &\vdots \\ &\leqslant \|I-BA\|^m \, \|x^{(0)} - x\| \end{aligned}$$

因为我们已假定 $\|I-AB\|<1$，所以对任何 $x^{(0)}$，当 $m\to\infty$时误差 $\|x^{(m)}-x\|$ 收敛于 0.

4.5.2 均衡化

在极端临界情况求解线性方程组时，许多细化可加到4.3节所述的分解和求解过程中去. 我们将简明地讨论5种这样的方法：

1. 用行均衡化预条件.
2. 用列均衡化预条件.
3. 全主元.
4. 在消元过程的每个主步中预条件或尺度化.
5. 在结束时迭代改进.

行均衡化是用每行中绝对值最大的元素除系数阵每行的过程，即，对 $1\leqslant i\leqslant n$，用 $r_i=1/\max\limits_{1\leqslant j\leqslant n}|a_{ij}|$ 乘第 i 行. 之后，新元素 $\tilde{a}_{ij}$ 将满足 $\max\limits_{1\leqslant j\leqslant n}|\tilde{a}_{ij}|=1$，$1\leqslant i\leqslant n$. 在二进制计算机的数值实践中，因子 r_i 取尽可能接近于 $1/\max\limits_{1\leqslant j\leqslant n}|a_{ij}|$ 的形如 2^m 的一个数. 这样做就是为了避免引入额外的舍入误差. 因为我们方程组中第 i 个方程或用 r_i 乘 b_i 得到.

$$\sum_{j=1}^{n}a_{ij}x_j=b_i \quad 或 \quad \sum_{j=1}^{n}(r_ia_{ij})x_j=r_ib_i$$

因此，数 r_i 在分解过程期间应该存放起来使得它们可在求解过程中使用. 按矩阵-向量记号，行均衡化可写成

$$(RA)x=(Rb)$$

其中 $R=\text{diag}(r_i)$.

列均衡化除了我们处理列外是类似的：对 $1\leqslant j\leqslant n$，我们用 $c_j=1/\max\limits_{1\leqslant i\leqslant n}|a_{ij}|$ 乘第 j 列. [203] 我们还是宁肯取 c_j 是一个尽可能接近 $1/\max\limits_{1\leqslant i\leqslant n}|a_{ij}|$ 的形如 2^m 的一个数. 我们的原方程为

$$\sum_{j=1}^{n}a_{ij}x_j=b_i \qquad (1\leqslant i\leqslant n)$$

现在可写成

$$\sum_{j=1}^{n}(c_ja_{ij})\left(\frac{x_j}{c_j}\right)=b_i \qquad (1\leqslant i\leqslant n)$$

在求解阶段之后，我们已算出分量为 x_j/c_j 的近似解. 为得到 x_j，这些解必须用 c_j 相乘. 因此，c_1，c_2，…，c_n 也应该存放起来. 利用矩阵，我们把列均衡化写成

$$(AC)(C^{-1}x)=b$$

其中 $C=\text{diag}(c_j)$.

若已对方程组执行了行和列均衡化，则全主元方法在起始步就可安全地简化为搜索矩阵中(数量上)最大元. 这个元素同时确定第1个主行和通过消元引入0的第1列. 因此，如果我们打算不按自然次序1，2，…，n 而按这个更准确的选主元方法所确定的次序处理列的话，将需要两个置换数组，一个列出相继的主元素的行数，而另一个列出相应的列数. (参见习题4.3.5和计算机习题4.3.1.)

在我们列举的改善中第4个方法是预条件或尺度化. 它为分解阶段提供了一个更为合理的结构，因为算法中的每一步除了应用于较小的矩阵外都像第1步那样做.

我们列举的第 5 个方法迭代改进已在前面作了讨论.

把行和列均衡化的值作为**预条件**这个过程是有点引起争论的. 我们看到行均衡化后接着用非行尺度主元的高斯消元法实际上和行尺度主元的高斯消元法相同. 因此，一个合理的方法是开始用列均衡化接着用行尺度主元的高斯消元法.

下面的例子指出行-列均衡化

$$\begin{bmatrix} 1 & 10^8 \\ 2 & 0 \end{bmatrix} \to \begin{bmatrix} 10^{-8} & 1 \\ 1 & 0 \end{bmatrix} \to \begin{bmatrix} 10^{-8} & 1 \\ 1 & 0 \end{bmatrix}$$

和列-行均衡化

$$\begin{bmatrix} 1 & 10^8 \\ 2 & 0 \end{bmatrix} \to \begin{bmatrix} \frac{1}{2} & 1 \\ 1 & 0 \end{bmatrix} \to \begin{bmatrix} \frac{1}{2} & 1 \\ 1 & 0 \end{bmatrix}$$

之间的差别. 注意，行-列均衡化的结果是用对角阵，譬如说 R 和 B，对方程组前-尺度和后-尺度化为

204

$$((RA)B)(B^{-1}x) = (Rb)$$

而列-行均衡化是对某些对角阵 C 和 S，相应地有

$$(S(AC))(C^{-1}x) = (Sb)$$

习题 4.5

1. 证明 $n\times n$ 可逆矩阵的集合是全体 $n\times n$ 矩阵集合中的一个开集. 因此，若 A 可逆，则存在一个正的 ϵ 使得每个满足 $\|A-B\|<\epsilon$ 的矩阵 B 也可逆.
2. 证明：若 A 可逆且 $\|B-A\|<\|A^{-1}\|^{-1}$，则 B 可逆.
3. 证明：若 $\|A\|<1$，则
$$\|(I-A)^{-1}\| \geqslant \frac{1}{1+\|A\|}$$
4. 证明：若 A 可逆且 $\|A-B\|<\|A^{-1}\|^{-1}$，则
$$\|A^{-1}-B^{-1}\| \leqslant \|A^{-1}\| \frac{\|I-A^{-1}B\|}{1-\|I-A^{-1}B\|}$$
5. 证明：若 $\|AB-I\|<1$，并用 E 表示 $AB-I$，则
$$A^{-1} = B - BE + BE^2 - BE^3 + \cdots$$
6. 证明：若 A 可逆，则对任何 B,
$$\|B-A^{-1}\| \geqslant \frac{\|I-AB\|}{\|A\|}$$
7. 证明或否定：若 $1=\|A\|>\|B\|$，则 $A-B$ 可逆.
8. 证明：若 $\|A\|<1$，则
$$(I+A)^{-1} = I - A + A^2 - A^3 + \cdots$$
9. 证明或否定：若 $\|AB-I\|<1$，则 $\|BA-I\|<1$.
10. 证明：若 $\|A\|<1$，则 $\|(I+A)^{-1}\|\leqslant(1-\|A\|)^{-1}$.
11. 证明：若 A 可逆且 $\|B-A\|<\|A^{-1}\|^{-1}$，则
$$B^{-1} = A^{-1}\sum_{k=0}^{\infty}(I-BA^{-1})^k$$

12. 对任何 $n\times n$ 矩阵 A，证明

$$A^m = I-(I-A)\sum_{k=0}^{m-1}A^k$$

13. 评论下面每个 $n\times n$ 矩阵是可逆的“证明”：给定 A，选择一个向量范数当使用相应的矩阵范数时使得 $\|I-A\|<1$. 然后应用与诺伊曼级数有关的定理.

205

14. 证明：若 $\inf_{\lambda\in\mathbb{R}}\|I-\lambda A\|<1$，则 A 可逆.
15. 证明可逆的 $n\times n$ 矩阵构成全体 $n\times n$ 矩阵集合的一个稠密集. 这意指：若 A 是 $n\times n$ 矩阵且 $\varepsilon>0$，则存在一个可逆阵 B 使 $\|A-B\|<\varepsilon$.（见习题 4.5.1.）
16. 证明若 $\|AB-I\|=\varepsilon<1$，则

$$\|A^{-1}-B\|\leqslant\|B\|\left(\frac{\varepsilon}{1-\varepsilon}\right)$$

17. 证明运算 $x\mapsto Ax$ 是连续的. 即对固定的 A，指出若一个序列$[x^{(k)}]$收敛于 x，则$[Ax^{(k)}]$收敛于 Ax.
18. 证明：若 E 是 $n\times n$ 矩阵，$\|E\|$ 充分小，则

$$\|(I-E)^{-1}-(I+E)\|\leqslant 3\|E\|^2$$

试问 $\|E\|$ 必须多小？
19. 证明：若 A 可逆，则

$$\|Ax\|\geqslant\|x\|\,\|A^{-1}\|^{-1}$$

20. 考虑区间$[0,1]$上定义的全体连续函数组成的向量空间 V. V 上两个重要的范数是

$$\|x\|_\infty=\max_{0\leqslant t\leqslant 1}|x(t)| \qquad \|x\|_1=\int_0^1|x(t)|\,\mathrm{d}t$$

指出函数 $x_n(t)=t^n$ 的序列具有性质 $\|x_n\|_\infty=1$ 且当 $n\to\infty$时，$\|x_n\|_1\to 0$. 从而，这些范数导致不同的收敛性概念.
21. 证明：若 $\|AB-I\|<1$，则在 $A(2B-BAB)$比较接近于 I 的意义下，$2B-BAB$ 与 B 相比是 A 的一个较好的近似逆.
22. 设 $B_k=\sum_{j=0}^{k}A^j$. 说明序列$[B_k]$可用公式 $B_0=I, B_{k+1}=I+AB_k$ 递归地计算.
23. 假设 $\|I-\alpha A\|<1$，α 是某个已知的纯量，给出一个表示 A^{-1}的级数.
24. 在赋范线性空间中，证明：若向量序列收敛，则它必定满足柯西准则.
25. 证明：若 A 病态，则在与 A 的距离 $\|A\|_2/\kappa(A)$之内存在一个奇异阵.
26. 对小的 $\delta>0$，考察线性方程组$\begin{bmatrix}1 & 2\\ 1+\delta & 2\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}3\\ 3+\delta\end{bmatrix}$.

    a. 利用近似解 $\tilde{x}=(3,0)^T$，比较残差向量的无穷范数和误差向量的无穷范数. 你能获得什么结果？

    b. 求条件数 $\kappa_\infty(A)$. 当 $\delta\to 0$ 时，情况怎样？

    c. 基于近似解 $\tilde{x}$，执行一步迭代改进.

27. 证明：若 $\|I-AB\|<1$，则 BA 可逆.（这里 A 和 B 为方阵. 当它们不是方阵时，情况怎样？）

206

28. 证明：若对某个 c 和某个整数 $n\geqslant 1$，$\|I-cA^n\|<1$，则 A 可逆.
29. 证明：若存在一个没有常数项的多项式 p 使得

$$\|I-p(A)\|<1$$

则 A 可逆. 推广上述结论使得 p 中系数可能是矩阵.
30. 证明：若 p 是有常数项 c_0 的多项式且 $|c_0|+\|I-p(A)\|<1$，则 A 可逆.

计算机习题 4.5

1. 本题的目的是编写和检验本书中给出的算法的改善过程．组成改善的条件是：

   ⅰ．列均衡化．

   ⅱ．行均衡化．

   ⅲ．全主元．

   ⅳ．包括在结束时二步迭代细化．或者编写一个独立的小程序求 A，x 和 b 并改进 x 二次(或 m 次)．这里，每个残差 $b_i-\sum_{j=1}^{n}a_{ij}x_j$ 应该用双精度计算然后舍入到单精度．注意，为了计算残差，原矩阵 A 和向量 b 必须存储起来．

   通过解 $Ax=b$ 来测试你的代码，这里 $n=10$ 且

$$\begin{cases} a_{ij}=(i/11)^j & (1\leqslant i,j\leqslant n) \\ b_i=i[1-(i/11)^{10}]/(33-3i) & (1\leqslant i\leqslant n) \end{cases}$$

   第 2 个测试，设 $n=4$，$a_{ij}=1/(2n-i-j+1)$ 且 $b_i=\sum_{j=1}^{n}a_{ij}$．解应该是 $x=(1,1,1,1)^{\mathrm{T}}$．第 3 个测试，用 $n=10$ 重复前面的测试．包括足够多的打印语句来说明许多细节——特别是初始解和用迭代改进后得到的解．

2. 利用 3×3 的测试矩阵 A 和前两个问题中的方法计算 $B=\sum_{j=0}^{20}A^j$ 并看是否 $(I-A)B\approx I$．对某个确信 $\|A\|<1$ 的从属矩阵范数．

3. 解下列线性方程组，并应用三步迭代改进．在每次迭代后打印 r，e，x．

$$\begin{bmatrix} 60 & 30 & 20 \\ 30 & 20 & 15 \\ 20 & 15 & 12 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}=\begin{bmatrix} 110 \\ 65 \\ 47 \end{bmatrix}$$

4.6 用迭代法解方程组

高斯算法及其变形称为求解矩阵问题 $Ax=b$ 的**直接法**．如没有舍入误差，它们通过有限步操作，产生完全精确的解 x．

相反，**间接法**产生一个理想地收敛于解的向量序列．当得到的近似解具有某种特定的精确度或在一定的迭代次数之后计算就停止．间接法本质上几乎总是**迭代的**：反复地应用一个简单的操作生成前面所提到的序列． 207

对含有成千个方程的大型线性方程组，迭代法从计算速度和计算机存储方面来看具有超过直接法的决定性优点．有时，当精确度要求不严格的话，适当的迭代次数就足以产生一个可接受的解．对**稀疏方程组**(其中 A 的大部分元素为 0)，迭代法通常是十分有效的．在稀疏问题中，A 的非零元有时以稀疏存储格式存放；在其他情况，根本就不需要存储 A！后面的情况在偏微分方程数值解中是很普遍的．此时，A 的每行根据需要可以生成，但使用后不必保留．迭代法的另一个优点是它们通常是稳定的．并且当继续操作时它们实际上会抑制误差(由于舍入或较小的失误)．

为传达一般的想法，我们描述两个基本的迭代法．

例 1 考察线性方程组

$$\begin{bmatrix} 7 & -6 \\ -8 & 9 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}=\begin{bmatrix} 3 \\ -4 \end{bmatrix}$$

试问如何用迭代过程求它的解?

解 对第 i 个未知数求解第 i 个方程的直接过程如下:

$$x_1^{(k)} = \frac{6}{7}x_2^{(k-1)} + \frac{3}{7}$$

$$x_2^{(k)} = \frac{8}{9}x_1^{(k-1)} - \frac{4}{9}$$

这称为**雅可比方法**或**雅可比迭代**. 最初,我们选择 $x_1^{(0)}$ 和 $x_2^{(0)}$ 作为最合用的解的猜测,或简单地取它们为 0. 然后上面的等式生成我们所希望的改进值 $x_1^{(1)}$ 和 $x_2^{(1)}$. 这个过程重复预定的次数或者直到发现向量$(x_1^{(k)}, x_2^{(k)})^T$ 中已达到一定的精确度. 下面是此例中雅可比方法迭代选择的若干值:

k	$x_1^{(k)}$	$x_2^{(k)}$
0	0.000 00	0.000 00
10	0.148 65	−0.198 20
20	0.186 82	−0.249 09
30	0.196 62	−0.262 15
40	0.199 13	−0.265 51
50	0.199 78	−0.266 37

208

显然这个迭代过程可加以修正使 $x_1^{(k)}$ 的最新的值能直接用于第 2 个等式中. 由此得到的方法称为**高斯-赛德尔方法**或**高斯-赛德尔迭代**. 它的等式是

$$x_1^{(k)} = \frac{6}{7}x_2^{(k-1)} + \frac{3}{7}$$

$$x_2^{(k)} = \frac{8}{9}x_1^{(k)} - \frac{4}{9}$$

由高斯-赛德尔方法输出的若干值如下:

k	$x_1^{(k)}$	$x_2^{(k)}$
0	0.000 00	0.000 00
10	0.219 78	−0.249 09
20	0.201 30	−0.265 31
30	0.200 09	−0.266 59
40	0.200 01	−0.266 66
50	0.200 00	−0.266 67

雅可比迭代和高斯-赛德尔迭代似乎收敛于相同的极限,而后者收敛较快. 注意,与直接法相反,我们得到的解的精度依赖于迭代过程何时停止. ■

4.6.1 基本概念

下面我们在更一般的数学背景下考察迭代法. 求解方程组

$$Ax = b \tag{1}$$

的一般类型迭代过程可以描述如下：预先指定某个称为**分裂矩阵**的矩阵 Q，并且把原问题改写成下面等价的形式

$$Qx = (Q-A)x+b \tag{2}$$

(2)式提出一个写成

$$Qx^{(k)} = (Q-A)x^{(k-1)}+b \qquad (k \geqslant 1) \tag{3}$$

的迭代过程. 初始向量 $x^{(0)}$ 可以是任意的；若可利用一个好的解的猜测，则应该把它作为 $x^{(0)}$. 如果(3)式中的迭代法对任意的初始向量 $x^{(0)}$ 收敛，我们就说迭代法收敛. 可以从(3)式中算出向量序列 $x^{(1)}$，$x^{(2)}$，…，而我们的目标是选取 Q 使下列两个条件满足：

1. 序列$[x^{(k)}]$容易计算.
2. 序列$[x^{(k)}]$迅速收敛于解.

在本节中，我们将看到如果容易求解 $Qx^{(k)}=y$ 且当 Q^{-1} 逼近 A^{-1} 时就可同时得到这两个条件.

209

我们注意到，若序列$[x^{(k)}]$收敛于一个向量 x，则 x 自然而然地是一个解. 当然，若我们仅在(3)式中取极限并利用代数运算的连续性，则结果为

$$Qx = (Q-A)x+b \tag{4}$$

这意味着 $Ax=b$.

为保证方程(1)对任意向量 b 有解，我们假定 A 非奇异. 同样假定 Q 非奇异使得(3)式可解未知向量 $x^{(k)}$. 在作了这些假定后，我们就可以使用下式：

$$x^{(k)} = (I-Q^{-1}A)x^{(k-1)}+Q^{-1}b \tag{5}$$

来作理论上的分析. 应当强调(5)式对此分析是方便的，但是数值计算 $x^{(k)}$ 几乎总是通过解(3)式得到而不用 Q^{-1}.

观察到实际的解 x 满足方程

$$x = (I-Q^{-1}A)x+Q^{-1}b \tag{6}$$

因此，x 是映射

$$x \mapsto (I-Q^{-1}A)x+Q^{-1}b \tag{7}$$

的一个不动点.

从(5)式减去(6)式中的项，得到

$$x^{(k)}-x = (I-Q^{-1}A)(x^{(k-1)}-x) \tag{8}$$

现在选取任何适宜的向量范数及其从属矩阵范数. 我们从(8)式得到

$$\| x^{(k)}-x \| \leqslant \| I-Q^{-1}A \| \, \| x^{(k-1)}-x \| \tag{9}$$

反复做这步，我们最终得出不等式

$$\| x^{(k)}-x \| \leqslant \| I-Q^{-1}A \|^{k} \, \| x^{(0)}-x \| \tag{10}$$

因此，当 $\| I-Q^{-1}A \| <1$ 时，我们立即得到对任意的 $x^{(0)}$，

$$\lim_{k\to\infty} \| x^{(k)}-x \| = 0 \tag{11}$$

注意到(根据 4.5 节中的定理 1)假设 $\| I-Q^{-1}A \| <1$ 可推出 $Q^{-1}A$ 和 A 的可逆性. 因此，我们有下面的定理.

定理1(迭代法收敛性定理) 若对某个从属矩阵范数 $\| I-Q^{-1}A \| < 1$，则由(3)式产生的序列对任意的初始向量 $x^{(0)}$ 收敛于 $Ax=b$ 的解.

若范数 $\delta \equiv \| I-Q^{-1}A \|$ 小于1，则当 $\| x^{(k)}-x^{(k-1)} \|$ 微小时，停止迭代过程是安全的. 甚至，我们可以证明(习题4.6.33)

210

$$\| x^{(k)}-x \| \leqslant \frac{\delta}{1-\delta} \| x^{(k)}-x^{(k-1)} \|$$

4.6.2 理查森方法

作为这些概念的一个说明，我们考察**理查森方法**，在这个方法中 Q 被选为单位阵. 此时(3)式变成

$$x^{(k)} = (I-A)x^{(k-1)} + b = x^{(k-1)} + r^{(k-1)} \tag{12}$$

其中 $r^{(k-1)}$ 是残差向量，$r^{(k-1)}=b-Ax^{(k-1)}$. 按定理1，如果对某个从属矩阵范数 $\| I-A \| < 1$ (见习题4.6.2～4.6.3，两类矩阵具有这样的性质)，理查森迭代(极限)会产生 $Ax=b$ 的一个解.

执行理查森迭代的算法如下：

```
input n, (a_ij), (b_i), (x_i), M
for k=1 to M do
    for i=1 to n do
        r_i ← b_i − Σ_{j=1}^{n} a_ij x_j
    end do
    for i=1 to n do
        x_i ← x_i + r_i
    end do
end do
output k, (x_i), (r_i)
```

例2 对下列问题利用理查森迭代法计算100次迭代，初始向量用 $x=(0,\ 0,\ 0)^T$.

$$\begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{3} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \frac{11}{18} \\ \frac{11}{18} \\ \frac{11}{18} \end{bmatrix}$$

解 写出基于上面算法的计算机程序. 由这个程序产生的一些迭代列举如下

$$\begin{aligned} x^{(0)} &= (0.000\,00,\quad 0.000\,00,\quad 0.000\,00)^T \\ x^{(1)} &= (0.611\,11,\quad 0.611\,11,\quad 0.611\,11)^T \\ &\vdots \\ x^{(10)} &= (0.279\,50,\quad 0.279\,50,\quad 0.279\,50)^T \\ &\vdots \end{aligned}$$

$$x^{(40)}=(0.333\,11,\quad 0.333\,11,\quad 0.333\,11)^{\mathrm{T}}$$
$$\vdots$$
$$x^{(80)}=(0.333\,33,\quad 0.333\,33,\quad 0.333\,33)^{\mathrm{T}}$$

■ 211

4.6.3 雅可比方法

我们基本理论的另一个说明是由**雅可比迭代**所提供的，在这方法中 Q 是对角阵，其对角元与矩阵 $A=(a_{ij})$ 中的那些对角元相同．此时，$Q^{-1}A$ 的一般元素是 a_{ij}/a_{ii}．这个矩阵的对角元都是 1，因此

$$\| I-Q^{-1}A\|_{\infty}=\max_{1\leqslant i\leqslant n}\sum_{\substack{j=1\\j\neq i}}^{n}|a_{ij}/a_{ii}| \tag{13}$$

定理 2（雅可比方法的收敛性定理） 若 A 对角占优，则对任意的初始向量，雅可比迭代产生的序列收敛于 $Ax=b$ 的解．

证明 对角占优意指

$$|a_{ii}|>\sum_{\substack{j=1\\j\neq i}}^{n}|a_{ij}| \qquad (1\leqslant i\leqslant n)$$

从(13)式，我们得到

$$\| I-Q^{-1}A\|_{\infty}<1$$

由定理 1，雅可比迭代收敛． ■

雅可比方法的算法如下：

```
input n, (a_ij), (b_i), (x_i), M
for k=1 to M do
    for i=1 to n do
        u_i ← (b_i − Σ_{j=1, j≠i}^{n} a_ij x_j) / a_ii
    end do
    for i=1 to n do
        x_i ← u_i
    end do
    output k, (x_i)
end do
```

这个算法和本节中其他算法如果在迭代开始以前先执行所有的除法即可使其更为有效．因此，我们可用下列这些运算开始计算：

```
for i=1 to n do
    d=1/a_ii
    b_i ← d b_i
    for j=1 to n do
        a_ij = d a_ij
    end do
end do
```

于是，u_i 的替换语句就变成

$$u_i \leftarrow b_i - \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}x_j$$

说明这点的另一种方式是原方程组 $Ax=b$ 用

$$D^{-1}Ax = D^{-1}b$$

替换，这里 $D=\mathrm{diag}(a_{ii})$. 另一方面，倘若 A 有正对角元，则可用其他一些尺度化预先处理方程组，从而可避免除法. 例如，双边尺度化

$$(D^{-1/2}AD^{-1/2})(D^{1/2}x) = (D^{-1/2}b)$$

其中 $D^{1/2}=\mathrm{diag}(\sqrt{a_{ii}})$. 注意：若 A 对称，则这个尺度化保持对称性. 在许多迭代法中，类似这样的简单预备工作会对效率或收敛速度带来很大的帮助.

4.6.4 分析

下面我们的工作是讨论任意线性迭代过程的某些理论. 我们考察用下列形式的式子所定义的过程

$$x^{(k)} = Gx^{(k-1)} + c \tag{14}$$

其中 G 是一个预先指定的 $n\times n$ 矩阵，c 是$\mathbb{R}^n$中预先指定的向量. 注意，(3)式定义的迭代被包括在我们可能对(14)式要讨论的所有一般理论中；即，我们可取 $G=I-Q^{-1}A$，$c=Q^{-1}b$. 我们希望对 G 求一个充要条件，使得(14)式的迭代对任意的初始向量收敛. 当然首先必须做一些准备工作.

矩阵 A 的**特征值**是使矩阵 $A-\lambda I$ 不可逆的复数 λ. 这些数是 A 的**特征方程**：

$$\det(A-\lambda I) = 0$$

的根.（读者可以提前看看 5.1 节，在那里我们进一步讨论了这些概念.）A 的**谱半径**定义为

212 ~ 213

$$\rho(A) = \max\{|\lambda|;\det(A-\lambda I) = 0\}$$

因此，在包含 A 的全部特征值的复平面上以 0 为中心的圆中，$\rho(A)$是最小的圆半径数.

如果存在一个非奇异阵 S 使得 $S^{-1}AS=B$，那么称 A **相似于**矩阵 B. 由此可得相似矩阵有相同的特征值. 进而，容易看出三角阵的特征值是位于其对角线上的元素.

定理 3（相似上三角阵定理） 每一个方阵相似于非对角元是任意小的（可能是复的）上三角阵.

证明 设 A 是 $n\times n$ 矩阵. 我们借用熟知的 5.2 节舒尔定理的结果. 这个定理表明 A 相似于一个上三角阵 $T=(t_{ij})$，它可能是复的. 现设 $0<\varepsilon<1$，并设 $D=\mathrm{diag}(\varepsilon,\varepsilon^2,\cdots,\varepsilon^n)$. 经初等计算，$D^{-1}TD$ 的一般元素是 $t_{ij}\varepsilon^{j-i}$. 因为 T 是上三角阵，所以 $D^{-1}TD$ 对角线下面的元素为 0，而由于 $j>i$ 且 $\varepsilon<1$，因此对角线上面的元素满足

$$|t_{ij}\varepsilon^{j-i}| \leqslant \varepsilon|t_{ij}|$$

通过变小 ε，这个上界可达到如我们所希望的那样小. ■

定理 4(谱半径定理)　谱半径函数满足等式

$$\rho(A)=\inf_{\|\cdot\|}\|A\|$$

其中下确界取遍所有从属矩阵范数.

证明　容易证明 $\rho(A)\leqslant\inf\limits_{\|\cdot\|}\|A\|$. 为此设 λ 是 A 的任意特征值. 选择一个对应于 λ 的非零特征向量 x. 则对任意的向量范数及其从属矩阵范数，我们有

$$|\lambda|\ \|x\|=\|\lambda x\|=\|Ax\|\leqslant\|A\|\ \|x\|$$

因此 $|\lambda|\leqslant\|A\|$. 由此可得 $\rho(A)\leqslant\|A\|$. 取下确界，我们就有 $\rho(A)\leqslant\inf\limits_{\|\cdot\|}\|A\|$.

对反向不等式，我们利用定理 3. 它断言，对任意 $\varepsilon>0$，存在一个非奇异阵 S 使得 $S^{-1}AS=D+T$，这里 D 是对角阵而 T 是严格上三角阵，并且 $\|T\|_\infty\leqslant\varepsilon$. 于是，我们有

$$\|S^{-1}AS\|_\infty=\|D+T\|_\infty\leqslant\|D\|_\infty+\|T\|_\infty$$

因为 D 在其对角线上有 A 的特征值，由此可得

$$\|D\|_\infty=\max_{1\leqslant i\leqslant n}|\lambda_i|=\rho(A)$$

214

因此，我们有

$$\|S^{-1}AS\|_\infty\leqslant\rho(A)+\varepsilon$$

借助于习题 4.6.6，我们知道由

$$\|A\|_\infty'\equiv\|S^{-1}AS\|_\infty$$

定义的函数 $\|\cdot\|_\infty'$ 是一个从属矩阵范数. 因此 $\|A\|_\infty'\leqslant\rho(A)+\varepsilon$，由取遍所有从属矩阵范数的下确界，得到

$$\inf_{\|\cdot\|}\|A\|\leqslant\rho(A)+\varepsilon$$

因为 ε 是任意的，所以 $\inf\limits_{\|\cdot\|}\|A\|\leqslant\rho(A)$. ■

定理 4 告诉我们，任何矩阵 A 的谱半径小于其范数(任意从属矩阵范数)值，而且还存在一个从属矩阵范数，其值任意接近谱半径.

现在我们给出与迭代矩阵 G 对应的迭代法收敛的充要条件.

定理 5(迭代法收敛性的充要条件定理)　为使迭代公式

$$x^{(k)}=Gx^{(k-1)}+c$$

对任意的初始向量 $x^{(0)}$ 产生一个收敛于 $(I-G)^{-1}c$ 的序列，它的充要条件是 G 的谱半径小于 1.

证明　假如 $\rho(G)<1$. 由定理 4，存在一个从属矩阵范数使得 $\|G\|<1$. 我们记

$$x^{(1)}=Gx^{(0)}+c$$
$$x^{(2)}=G^2x^{(0)}+Gc+c$$
$$x^{(3)}=G^3x^{(0)}+G^2c+Gc+c$$

一般公式是

$$x^{(k)}=G^kx^{(0)}+\sum_{j=0}^{k-1}G^jc \tag{15}$$

利用形成矩阵范数的向量范数，我们有

$$\|G^kx^{(0)}\|\leqslant\|G^k\|\ \|x^{(0)}\|\leqslant\|G\|^k\|x^{(0)}\|\to 0,\quad 当\ k\to\infty$$

215

由 4.5 节定理 1，我们有

$$\sum_{j=0}^{\infty} G^j c = (I-G)^{-1}c$$

于是，在(15)式中让 $k\to\infty$，得到

$$\lim_{k\to\infty} x^{(k)} = (I-G)^{-1}c$$

反之，假如 $\rho(G)\geqslant 1$. 选择 u 和 λ 使得

$$Gu = \lambda u \qquad |\lambda|\geqslant 1 \qquad u\neq 0$$

设 $c=u$ 和 $x^{(0)}=0$. 由(15)式，$x^{(k)}=\sum_{j=0}^{k-1} G^j u=\sum_{j=0}^{k-1}\lambda^j u$. 若 $\lambda=1$，则 $u^{(k)}=ku$ 并且当 $k\to\infty$时，它发散. 若 $\lambda\neq 1$，则 $x^{(k)}=(\lambda^k-1)(\lambda-1)^{-1}u$，并且因为$\lim_{k\to\infty}\lambda^k$ 不存在，所以它也发散. ■

推论 1(迭代法收敛性推论) 迭代公式(3)，即 $Qx^{(k)}=(Q-A)x^{(k-1)}+b$，当 $\rho(I-Q^{-1}A)<1$ 时，对任意的 $x^{(0)}$，将产生一个收敛于 $Ax=b$ 解的序列.

4.6.5 高斯-赛德尔方法

让我们更仔细地考察**高斯-赛德尔**迭代. 设 Q 是由 A 的下三角部分包括对角线所定义的.

定理 6(高斯-赛德尔方法收敛性定理) 若 A 对角占优，则对任何初始向量高斯-赛德尔方法收敛.

证明 由推论 1，证明

$$\rho(I-Q^{-1}A)<1$$

就足够了. 为此，设 λ 是 $I-Q^{-1}A$ 的任意特征值，x 是对应的特征向量. 不失一般性，我们假定 $\|x\|_\infty=1$. 现在有

$$(I-Q^{-1}A)x=\lambda x \quad 或 \quad Qx-Ax=\lambda Qx$$

因为 Q 是 A 的下三角部分，包括其对角线，所以

$$-\sum_{j=i+1}^{n} a_{ij}x_j = \lambda\sum_{j=1}^{i} a_{ij}x_j \qquad (1\leqslant i\leqslant n)$$

对这个方程移项，得到

[216]

$$\lambda a_{ii}x_i = -\lambda\sum_{j=1}^{i-1} a_{ij}x_j - \sum_{j=i+1}^{n} a_{ij}x_j \qquad (1\leqslant i\leqslant n)$$

选择一个指标 i 使得对一切 j 有 $|x_i|=1\geqslant|x_j|$. 于是

$$|\lambda||a_{ii}|\leqslant|\lambda|\sum_{j=1}^{i-1}|a_{ij}|+\sum_{j=i+1}^{n}|a_{ij}|$$

求解 $|\lambda|$，并利用 A 的对角占优性，我们得到

$$|\lambda|\leqslant\left\{\sum_{j=i+1}^{n}|a_{ij}|\right\}\left\{|a_{ii}|-\sum_{j=1}^{i-1}|a_{ij}|\right\}^{-1}<1$$ ■

高斯-赛德尔迭代的算法如下：

```
input n, (a_ij), (b_i), (x_i), M
for k=1 to M do
    for i=1 to n do
```

```
        x_i ← (b_i − Σ_{j=1, j≠i}^{n} a_ij x_j) / a_ii
    end do
    output k, (x_i)
end do
```

注意：在高斯-赛德尔算法中，更新的 x_i 值直接替代旧的值，而在雅可比方法中，进行替代之前要先算出 x 向量的所有新分量. 在雅可比算法中，可以同时计算 x 的新分量(在伪代码中用 u_i 表示)，而在高斯-赛德尔方法中它们必须串行计算，因为计算新的 x_i 需要所有新的 x_1，x_2，…，x_{i-1}值. 由于这个差别，雅可比迭代在允许向量处理或并行处理的计算机上可能更为可取. 注意，在高斯-赛德尔算法中改进的效果也可以通过对方程组执行某些预备的工作得到. 事实上，关于雅可比迭代应用所作的注记在这里未改变.

例 3 考察方程组

$$\begin{bmatrix} 2 & -1 & 0 \\ 1 & 6 & -2 \\ 4 & -3 & 8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ -4 \\ 5 \end{bmatrix}$$

应用高斯-赛德尔迭代. 初始向量 $x^{(0)}=(0,\ 0,\ 0)^{\mathrm{T}}$.

解 借助前面的尺度化，$D^{-1}Ax=D^{-1}b$，这里 $D=\mathrm{diag}(A)$，方程组变为

$$\begin{bmatrix} 1 & -\frac{1}{2} & 0 \\ \frac{1}{6} & 1 & -\frac{1}{3} \\ \frac{1}{2} & -\frac{3}{8} & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -\frac{2}{3} \\ \frac{5}{8} \end{bmatrix}$$

217

我们把这个方程组记为 $Ax=b$. 在高斯-赛德尔算法中，Q 取 A 的下三角部分，包括对角线. 所确定的迭代公式是

$$Qx^{(k)} = (Q-A)x^{(k-1)} + b$$

或

$$\begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{6} & 1 & 0 \\ \frac{1}{2} & -\frac{3}{8} & 1 \end{bmatrix} \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ x_3^{(k)} \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1^{(k-1)} \\ x_2^{(k-1)} \\ x_3^{(k-1)} \end{bmatrix} + \begin{bmatrix} 1 \\ -\frac{2}{3} \\ \frac{5}{8} \end{bmatrix}$$

由此，解下三角方程组得到 $x^{(k)}$. 此例中有关的公式是

$$\begin{aligned} x_1^{(k)} &= \frac{1}{2}x_2^{(k-1)} + 1 \\ x_2^{(k)} &= -\frac{1}{6}x_1^{(k)} + \frac{1}{3}x_3^{(k-1)} - \frac{2}{3} \\ x_3^{(k)} &= -\frac{1}{2}x_1^{(k)} + \frac{3}{8}x_2^{(k)} + \frac{5}{8} \end{aligned}$$

通过计算得到下列迭代，其中 $x^{(13)}$ 是正确的：

$$x^{(1)} = (1.000\,000, -0.833\,333, -0.187\,500)^T$$
$$\vdots$$
$$x^{(5)} = (0.622\,836, -0.760\,042, 0.028\,566)^T$$
$$\vdots$$
$$x^{(10)} = (0.620\,001, -0.760\,003, 0.029\,998)^T$$
$$\vdots$$
$$x^{(13)} = (0.620\,000, -0.760\,000, 0.030\,000)^T$$

■

4.6.6 SOR 方法

下一个重要的迭代法例子是熟知的**逐次超松弛法**，通常缩写为 **SOR**. 因为我们希望能提供应用于复数域上矩阵和向量的 SOR 的一般理论，所以我们首先复习与这个背景有关的一些概念.

若 γ 是一个复数，则 γ 可写成形式 $\gamma=\alpha+\mathrm{i}\beta$，这里 α 和 β 是实数且 $\mathrm{i}^2=-1$. γ 的**共轭**定义为

$$\bar{\gamma} = \alpha - \mathrm{i}\beta$$

γ 的**大小**是 $|\gamma| = \sqrt{\alpha^2+\beta^2} = \sqrt{\gamma\bar{\gamma}}$. 用 $\mathbb{C}^n$ 表示复 n 维向量空间. 在这个空间中，用等式

218

$$\langle x,y\rangle = y^* x = \sum_{i=1}^{n} x_i \bar{y}_i$$

定义内积，这里 y^* 是由 $y^*=(\bar{y}_1, \bar{y}_2, \cdots, \bar{y}_n)$ 定义的行向量，称之为 y 的**共轭转置**. 容易看出

$$\langle x,x\rangle > 0 \qquad (若\ x \neq 0)$$
$$\langle x,\lambda y\rangle = \bar{\lambda}\langle x,y\rangle$$
$$\langle x,y\rangle = \overline{\langle y,x\rangle}$$

由此可得，对纯量 α 和 β 以及向量 x，y，z，有 $\langle \alpha x+\beta y, z\rangle = \alpha\langle x, z\rangle + \beta\langle y, z\rangle$ 以及 $\langle x, Ay\rangle = \langle A^* x, y\rangle$. x 的**欧几里得范数**是

$$\|x\|_2 = \sqrt{\langle x,x\rangle} = \sqrt{x^* x} = \left\{\sum_{i=1}^{n} |x_i|^2\right\}^{1/2}$$

矩阵 $A=(a_{ij})$，如果 $A^*=A$，就称 A 是**埃尔米特阵**. 这里 $A^*=(\bar{a}_{ji})$ 是 A 的**共轭转置**. 最后，若对一切 $x\neq 0$，$\langle Ax, x\rangle > 0$，则矩阵 A 称为**正定的**. 注意：若 A 是埃尔米特阵，则 $\langle Ax, y\rangle = \langle x, Ay\rangle$.

下面，当 A 是埃尔米特阵和正定阵时，我们提出一个含有 SOR 方法收敛性条件的定理.

定理 7（SOR 方法收敛性定理） 在 SOR 方法中，假如选分裂阵 Q 为 $\alpha D-C$，这里 α 是一个实参数，D 是任意埃尔米特正定阵而 C 是满足 $C+C^*=D-A$ 的任意矩阵. 若 A 是埃尔米特正定阵，Q 非奇异，且 $\alpha>1/2$，则对任意初始向量 SOR 迭代收敛.

证明 如前面证明中那样，我们设 $G=I-Q^{-1}A$，试图证明 G 的谱半径满足 $\rho(G)<1$. 设 λ 是 G 的一个特征值并设 x 是对应于 λ 的一个特征向量. 取 $y=(I-G)x$. 则很容易验证下列

等式：

$$y = x - Gx = x - \lambda x = Q^{-1}Ax \tag{16}$$

$$Q - A = (\alpha D - C) - (D - C - C^*) = \alpha D - D + C^* \tag{17}$$

利用(16)式，我们有

$$(\alpha D - C)y = Qy = Ax \tag{18}$$

利用(17)，(18)，(16)式我们得到

$$\begin{aligned}(\alpha D - D + C^*)y &= (Q - A)y = Ax - Ay = A(x - y)\\ &= A(x - Q^{-1}Ax) = AGx\end{aligned} \tag{19}$$

从(18)式和(19)式，我们有

$$\alpha\langle Dy, y\rangle - \langle Cy, y\rangle = \langle Ax, y\rangle \tag{20}$$

$$\alpha\langle y, Dy\rangle - \langle y, Dy\rangle + \langle y, C^* y\rangle = \langle y, AGx\rangle \tag{21}$$

219

将(20)式和(21)式相加，我们得到

$$2\alpha\langle Dy, y\rangle - \langle y, Dy\rangle = \langle Ax, y\rangle + \langle y, AGx\rangle \tag{22}$$

$$(2\alpha - 1)\langle Dy, y\rangle = \langle Ax, y\rangle + \langle y, AGx\rangle \tag{23}$$

其中我们利用了$\langle Dy, y\rangle = \langle y, Dy\rangle$，因为$D$是埃尔米特阵．因为$y = (1-\lambda)x$和$Gx = \lambda x$，所以由(23)式产生

$$\begin{aligned}(2\alpha - 1)\,|1 - \lambda|^2\langle Dx, x\rangle &= (1 - \bar{\lambda})\langle Ax, x\rangle + \bar{\lambda}(1 - \lambda)\langle x, Ax\rangle\\ &= (1 - |\lambda|^2)\langle Ax, x\rangle\end{aligned} \tag{24}$$

最后一个等式利用了A是埃尔米特阵．若$\lambda \neq 1$，则(24)式的左边为正的．因此，右边也必须为正的，且$|\lambda| < 1$．另一方面，若$\lambda = 1$，则由$y = (1-\lambda)x$知$y = 0$且从(18)式知$Ax = 0$．这与任何$x \neq 0$，$\langle Ax, x\rangle > 0$的条件矛盾．因此，$\rho(G) < 1$并且SOR收敛． ■

在SOR方法中，通常选取D为A的对角元而设$-C$是A的下三角部分，不包括对角线．然而，定理7未预先假定这个选择．还有，读者应该注意这样一个事实，就是在文献中参数α通常用$1/\omega$表示．于是$0 < \omega < 2$．在Young[1971]、Varga[1962]、Hageman and Young[1981]、Wachspress[1966]、Issacson and Keller[1966]，以及其他许多著作中都涉及如何选择w使得SOR迭代具有最快收敛性的问题．

4.6.7 迭代矩阵

假如A分块成

$$A = D - C_L - C_U$$

这里$D = \mathrm{diag}(A)$，C_L是$-A$的严格下三角部分而C_U是$-A$的严格上三角部分．另一种分块是类似的，但是用块分量．在偏微分方程的离散化中，第一种分块对应于单个网格点，而后者对应于诸如网格点线或网格点块那样的组合．

对本节提出的基本迭代法，我们总结关键的矩阵和方法如下：

理查森：

$$\begin{cases} Q = I \\ G = I - A \end{cases}$$

$$x^{(k)} = (I-A)x^{(k-1)} + b$$

雅可比：

$$\begin{cases} Q = D \\ G = D^{-1}(C_L + C_U) \end{cases}$$

[220]

$$Dx^{(k)} = (C_L + C_U)x^{(k-1)} + b$$

高斯-赛德尔：

$$\begin{cases} Q = D - C_L \\ G = (D - C_L)^{-1}C_U \end{cases}$$

$$(D - C_L)x^{(k)} = C_U x^{(k-1)} + b$$

SOR：

$$\begin{cases} Q = \omega^{-1}(D - \omega C_L) \\ G = (D - \omega C_L)^{-1}(\omega C_U + (1-\omega)D) \end{cases}$$

$$(D - \omega C_L)x^{(k)} = \omega(C_U x^{(k-1)} + b) + (1-\omega)Dx^{(k-1)}$$

SSOR：

$$\begin{cases} Q = (\omega(2-\omega))^{-1}(D - \omega C_L)D^{-1}(D - \omega C_U) \\ G = (D - \omega C_U)^{-1}(\omega C_L + (1-\omega)D)(D - \omega C_L)^{-1}(\omega C_U + (1-\omega)D) \end{cases}$$

$$(D - \omega C_L)x^{(k-1/2)} = \omega(C_U x^{(k-1)} + b) + (1-\omega)Dx^{(k-1)}$$

$$(D - \omega C_U)x^{(k)} = \omega(C_L x^{(k-1/2)} + b) + (1-\omega)Dx^{(k-1/2)}$$

这里我们已包含了另一个基本迭代法：**对称逐次超松弛(SSOR)法**. SSOR 的每个迭代都由下面两步组成，第一步是**向前 SOR** 迭代，以一定的次序计算未知量，第二步是**向后 SOR** 扫描，以反向次序求解它们. 为 SOR 和 SSOR 方法选择最佳松弛参数是一个解答相当复杂而又引人入胜的问题，我们在此不加讨论.

4.6.8 外推

下面我们将介绍一个称为**外推**的一般方法，它可用来改进线性迭代过程的收敛性质. 考察迭代公式

$$x^{(k)} = Gx^{(k-1)} + c \tag{25}$$

我们引进参数 $\gamma \neq 0$，并把方程(25)嵌入到下面单参数的迭代方程族中，

$$\begin{aligned} x^{(k)} &= \gamma(Gx^{(k-1)} + c) + (1-\gamma)x^{(k-1)} \\ &= G_\gamma x^{(k-1)} + \gamma c \end{aligned} \tag{26}$$

其中

$$G_\gamma = \gamma G + (1-\gamma)I$$

注意：当 $\gamma = 1$ 时，我们重新获得(25)式中原来的迭代.

若(26)中的迭代收敛于 x，则取极限，我们得到

[221]

$$x = \gamma(Gx + c) + (1-\gamma)x$$

或者，因为 $\gamma \neq 0$，可得到

$$x = Gx + c$$

记得迭代(25)的目标是产生方程 $x=Gx+c$ 的解. 若 $G=I-Q^{-1}A$ 及 $c=Q^{-1}b$，则这个方法对应于求解 $Ax=b$.

在试图确定参数 γ 的最优值之前，我们需要一个有关特征值的结果.

定理 8($p(A)$的特征值定理) 若 λ 是矩阵 A 的特征值，并且 p 是一个多项式，则 $p(\lambda)$是 $p(A)$的特征值.

证明 设 $Ax=\lambda x$，$x\neq 0$. 则 $A^2x=\lambda Ax=\lambda^2 x$. 由归纳法及对 $k=0$ 的单独验证，我们有

$$A^k x = \lambda^k x \ (k \geqslant 0)$$

因此，λ^k 是 A^k 的特征值. 对多项式 p，记 $p(z)=\sum_{k=0}^{m} c_k z^k$. 则

$$p(A)x = \sum_{k=0}^{m} c_k A^k x = \sum_{k=0}^{m} c_k \lambda^k x = p(\lambda)x$$

■

由定理 5，(26)式中的外推法收敛的充要条件是 $\rho(G_\gamma)<1$. 假如我们不能准确地知道 G 的特征值，而仅仅知道直线上包含 G 的全部特征值的一个区间$[a, b]$. 由定理 8，矩阵 $G_\gamma\equiv\gamma G+(1-\gamma)I$ 的特征值位于端点为 $\gamma a+1-\gamma$ 和 $\gamma b+1-\gamma$ 的区间中. 试问有可能选择 γ 使得 $\rho(G_\gamma)<1$ 吗?

用 $\Lambda(A)$表示任意矩阵 A 的特征值集合. 则

$$\rho(G_\gamma) = \max_{\lambda\in\Lambda(G_\gamma)} |\lambda| = \max_{\lambda\in\Lambda(G)} |\gamma\lambda+1-\gamma| \leqslant \max_{a\leqslant\lambda\leqslant b} |\gamma\lambda+1-\gamma| \tag{27}$$

我们将证明：若 $1\notin[a, b]$，则可选择 γ 使得 $\rho(G_\gamma)<1$.

定理 9(最佳外推参数定理) 若关于 G 的特征值仅有的可用信息是它们位于区间$[a, b]$中且 $1\notin[a, b]$，则 γ 的最好选择是 $2/(2-a-b)$. 对这个 γ 的值，$\rho(G_\gamma)\leqslant 1-|\gamma|d$，这里 d 是从 1 到$[a, b]$的距离.

证明 假定假设成立且 γ 已知. 因为 $1\notin[a, b]$，所以或者 $a>1$ 或者 $b<1$. 我们只给出第 2 种情况的证明，第 1 种情况的证明留作习题. 因为 $a\leqslant b<1$，由此可得 $\gamma>0$ 且 $d=1-b$. 因此，如前段所注明的那样，G_γ 的任意特征值 λ 满足不等式

$$\gamma a+1-\gamma \leqslant \lambda \leqslant \gamma b+1-\gamma \tag{28}$$

222

因此，

$$\lambda \leqslant \gamma b+1-\gamma = 1+\gamma(b-1) = 1-\gamma d$$

此外，

$$\lambda \geqslant \gamma a+1-\gamma = \gamma(a+b-2)+1+\gamma(1-b) = -1+\gamma d$$

因而，我们证得

$$-1+\gamma d \leqslant \lambda \leqslant 1-\gamma d$$

因此 $\rho(G_\gamma)\leqslant 1-\gamma d$. 为了看出我们选择的 γ 是最佳的，注意当 γ 递增时，(28)式中区间的左端点向左移动. 如果那个点碰巧是 G_γ 的一个特征值，那么 $\rho(G_\gamma)$就递增. 当 γ 递减时，类似可证.

■

应该看到刚才所讨论的外推过程也可应用于它们本身并不收敛的方法中. 而这一切所需要的是 G 的特征值是实的且位于不包含 1 的一个区间中.

若 A 是一个矩阵，其特征值 λ_1，λ_2，…，λ_n 全部是实的，我们定义

$$m(A)=\min_i \lambda_i \qquad M(A)=\max_i \lambda_i \tag{29}$$

因而，在定理 9 中，我们可设 $a=m(G)$，$b=M(G)$.

例 4 确定最佳外推理查森方法的谱半径.

解 在理查森迭代情况中，$Q=I$，$G=I-A$. 若 A 只有实特征值，则 G 的特征值也是实的. 由定理 8，我们有

$$M(G)=1-m(A) \qquad m(G)=1-M(A) \tag{30}$$

若 $m(A)>0$ 或 $M(A)<0$，则在理查森迭代中加速是可能的. 由定理 9 计算的最佳 γ 是

$$\gamma=2/[m(A)+M(A)]$$

由 $d=m(A)$，算得的谱半径是

$$\rho(G_\gamma)=[M(A)-m(A)]/[M(A)+m(A)]$$

■

例 5 确定尺度化方程组的最佳外推雅可比方法谱半径.

解 雅可比迭代可如习题 4.6.9 中指出的那样处理. 我们设

223

$$D=\operatorname{diag}(a_{11},a_{22},\cdots,a_{nn})$$

并应用理查森迭代于问题 $D^{-1}Ax=D^{-1}b$. 由例 4 的结果，我们得到：若 $m(D^{-1}A)>0$ 或 $M(D^{-1}A)<0$，则在理查森迭代中加速是可能的. 此外，如果使用最佳的 $\gamma=2/[m(D^{-1}A)+M(D^{-1}A)]$，那么迭代矩阵的谱半径是

$$\rho(G_\gamma)=[M(D^{-1}A)-m(D^{-1}A)]/[M(D^{-1}A)+m(D^{-1}A)]$$

■

4.6.9 切比雪夫加速

更一般类型的加速过程称为**切比雪夫加速**，也可应用于线性迭代算法. 如前面一样，先考察基本迭代法

$$x^{(k)}=Gx^{(k-1)}+c$$

假定问题的解是满足 $x=Gx+c$ 的向量 x. 在过程中的第 k 步，我们已算得向量 $x^{(1)}$，$x^{(2)}$，…，$x^{(k)}$，我们要问这些向量的某个线性组合是否会比 $x^{(k)}$ 更好地逼近于解. 我们假定 $a_0^{(k)}+a_1^{(k)}+\cdots+a_k^{(k)}=1$，并取

$$u^{(k)}=\sum_{i=0}^{k}a_i^{(k)}x^{(i)}$$

利用所熟悉的方法，我们得到

$$\begin{aligned}u^{(k)}-x&=\sum_{i=0}^{k}a_i^{(k)}(x^{(i)}-x)=\sum_{i=0}^{k}a_i^{(k)}G^i(x^{(0)}-x)\\&=P(G)(x^{(0)}-x)\end{aligned}$$

其中 P 是由 $P(z)=\sum_{i=0}^{k}a_i^{(k)}z^i$ 定义的多项式. 取范数，我们得到

$$\|u^{(k)}-x\|\leqslant\|P(G)\|\,\|x^{(0)}-x\|$$

这里可以使用任何向量范数及其从属矩阵范数. 从定理 4 知，取遍全体从属矩阵范数的 $\|P(G)\|$，其下确界是 $\rho(P(G))$；这个下确界就是应该使其为最小的值. 若 G 的特征值 μ_i 位

于复平面上某个有界集 S 中，则由定理 8，得

$$\rho(P(G)) = \max_{1\leqslant i\leqslant n} | P(\mu_i) | \leqslant \max_{z\in S} | P(z) |$$

应该极小化最后的表达式来选定多项式 P，并且服从约束 $\sum_{i=0}^{k} a_i=1$，这里 $P(1)=1$. 这是逼近论中的一个标准问题，在某些情况其显式解是已知的.

例如，若 S 是实直线上一个不包含 1 的区间$[a, b]$，则一个尺度化和位移的切比雪夫多项式可解决这个问题. 经典的切比雪夫多项式 $T_k(k\geqslant 1)$是极小化表达式 [224]

$$\max_{-1\leqslant z\leqslant 1} | T_k(z) |$$

服从首项系数为 2^{k-1}约束的唯一 k 次多项式. 这些多项式由下列公式递归地生成：

$$\begin{cases} T_0(z) = 1 \qquad T_1(z) = z \\ T_k(z) = 2zT_{k-1}(z) - T_{k-2}(z) \qquad (k \geqslant 2) \end{cases}$$

现假定 G 的特征值位于不含 1 的区间$[a, b]$中，譬如说 $b<1$. 我们对下列 min-max 问题感兴趣：

$$\min_{P_k(1)=1} \{ \max_{a\leqslant z\leqslant b} | P_k(z) | \}$$

这个极值问题的解包括在下列四个引理中. 这些结果涉及到上面和 6.1 节中所讨论的切比雪夫多项式 T_k. 用 Π_k 表示次数至多是 k 次的多项式集合.

引理 1（规范多项式第 1 引理） 设 $\beta \in \mathbb{R} \setminus (-1,1)$. 若 $p \in \Pi_k$ 且 $p(\beta) = 1$，则 $\| p \| \geqslant | \alpha |$，这里 $\alpha = 1/T_k(\beta)$ 且 $\| p \| = \max_{-1\leqslant t\leqslant 1} | p(t) |$.

证明 假如 p 满足假设但结论不成立. 设 $t_i=\cos(i\pi/k)$，$0\leqslant i\leqslant k$. 这些点是 T_k 的极值点. 当然，

$$T_k(t_i) = \cos(k \cos^{-1} t_i) = \cos i\pi = (-1)^i$$

设 $\sigma=\operatorname{sgn} \alpha$，则

$$\sigma(-1)^i[\alpha T_k(t_i) - p(t_i)] \geqslant | \alpha | - \| p \| > 0$$

这表明多项式 $\alpha T_k - p$ 在点 t_0，t_1，…，t_k 上交替取正值和负值. 因此，这个多项式在区间$(-1,1)$中至少有 k 个零点. 因为在点 β 它也为零，所以这个多项式至少共计有 $k+1$ 个零点. 因为 $\alpha T_k - p$ 的次数至多为 k 次，所以它必须为 0，矛盾. ■

引理 2（规范多项式第 2 引理） 设 $a<b<1$. 若 $p\in \Pi_k$ 且 $p(1)=1$，则 $\| p \| \geqslant 1/T_k(w(1))$. 这里

$$\| p \| = \max_{a\leqslant t\leqslant b} | p(t) | \qquad 且 \qquad w(t) = (2t-b-a)/(b-a)$$

若 $p=(T_k \circ w) / T_k(w(1))$，则不等式变成等式. [225]

证明 取 $\beta=w(1)$且 $p=q \circ w$，$q\in \Pi_k$. 注意 $1=p(1)=q(w(1))=q(\beta)$且 $\beta>1$. 由引理 1 立即可得 $\| q \|_{[-1,1]} \geqslant 1/T_k(\beta)$. 等价地，我们有 $\| p \|_{[a,b]} \geqslant 1/T_k(w(1))$. 若 $p=(T_k \circ w)/T_k(w(1))$，则显然 $p(1)=1$ 且 $\| p \|_{[a,b]} = \| T_k \|_{[-1,1]}/T_k(w(1))=1/T_k(w(1))$. ■

引理 3（多项式 P_k，递归关系引理） 多项式 $P_k=(T_k \circ w) / T_k(w(1))$可由递归关系：

$$\begin{cases} P_0(t) = 1 \quad P_1(t) = (2t - b - a)/(2 - b - a) \\ P_k(t) = \rho_k P_1(t) P_{k-1}(t) + (1 - \rho_k) P_{k-2}(t) \qquad (k \geqslant 2) \end{cases}$$

生成，其中系数 ρ_k 由下列等式得到

$$\rho_1 = 2 \qquad \rho_k = (1 - \alpha\rho_{k-1})^{-1} \qquad \alpha = [2w(1)]^{-2} \qquad (k \geqslant 2)$$

证明 定义 $\beta_k = T_k(w(1))$. 借助于关系 $T_k(t) = 2t\,T_{k-1}(t) - T_{k-2}(t)$ 和 $\beta_k P_k(t) = T_k(w(t))$，我们得到

$$\begin{aligned} P_k(t) = \beta_k^{-1} T_k(w(t)) &= \beta_k^{-1}[2w(t) T_{k-1}(w(t)) - T_{k-2}(w(t))] \\ &= 2\beta_k^{-1}\beta_{k-1} w(t) P_{k-1}(t) - \beta_k^{-1}\beta_{k-2} P_{k-2}(t) \\ &= 2\beta_k^{-1}\beta_{k-1} w(1) P_1(t) P_{k-1}(t) - \beta_k^{-1}\beta_{k-2} P_{k-2}(t) \end{aligned}$$

这里，我们利用容易验证的等式 $w(1)P_1(t) = w(t)$. 定义 $\rho_k = 2\beta_k^{-1}\beta_{k-1}\, w(1) = \alpha^{-1/2}\beta_k^{-1}\,\beta_{k-1}$ 是很方便的. 再利用切比雪夫多项式的递归关系，得到

$$\beta_k = T_k(w(1)) = 2w(1) T_{k-1}(w(1)) - T_{k-2}(w(1)) = \alpha^{-1/2}\beta_{k-1} - \beta_{k-2}$$

$$1 = 2w(1)\beta_k^{-1}\beta_{k-1} - \beta_k^{-1}\beta_{k-2} = \rho_k - \beta_k^{-1}\beta_{k-2}$$

于是，P_k 的递归关系可写成

$$P_k = \rho_k P_1 P_{k-1} + (1 - \rho_k) P_{k-2}$$

系数 ρ_k 满足等式

$$\begin{aligned} \rho_k = \alpha^{-1/2}\beta_k^{-1}\beta_{k-1} &= \alpha^{-1/2}\beta_{k-1}[\alpha^{-1/2}\beta_{k-1} - \beta_{k-2}]^{-1} \\ &= \alpha^{-1/2}[\alpha^{-1/2} - \beta_{k-1}^{-1}\beta_{k-2}]^{-1} \\ &= \alpha^{-1}\{\alpha^{-1} - \alpha^{-1/2}\beta_{k-1}^{-1}\beta_{k-2}\}^{-1} \\ &= \alpha^{-1}\{\alpha^{-1} - \rho_{k-1}\}^{-1} \\ &= (1 - \alpha\rho_{k-1})^{-1} \end{aligned}$$

226 ■

引理 4（切比雪夫加速，递归公式引理） 切比雪夫加速方法中的向量 $u^{(k)}$ 可从任意初始向量 $u^{(0)}$ 出发用下列公式递归计算：

$$\begin{cases} u^{(1)} = \gamma[Gu^{(0)} + c] + (1 - \gamma)u^{(0)} \qquad (\gamma = 2/(2 - b - a)) \\ u^{(k)} = \rho_k[\gamma(Gu^{(k-1)} + c) + (1 - \gamma)u^{(k-1)}] + (1 - \rho_k)u^{(k-2)} \qquad (k \geqslant 2) \end{cases}$$

证明 我们保持前面建立的所有记号. 特别地，

$$u^{(k)} = \sum_{i=0}^{k} a_i^{(k)} x^{(i)} \qquad P_k(t) = \sum_{i=0}^{k} a_i^{(k)} t^i$$

由此可得 $u^{(0)} = x^{(0)}$ 而 $u^{(1)}$ 由引理中的公式给出. 这留给读者验证. 对满足 $x = Gx + c$ 的 x，从前面证明的式子，我们有

$$\begin{aligned} u^{(k)} - x &= P_k(G)(u^{(0)} - x) \\ &= [\rho_k P_1(G) P_{k-1}(G) + (1 - \rho_k) P_{k-2}(G)](u^{(0)} - x) \\ &= \rho_k P_1(G)(u^{(k-1)} - x) + (1 - \rho_k)(u^{(k-2)} - x) \end{aligned}$$

这可写成下列形式

$$u^{(k)} = \rho_k P_1(G) u^{(k-1)} + (1 - \rho_k) u^{(k-2)} + \rho_k[I - P_1(G)]x$$

一个容易的计算显示这个式子的最后项等于 $\rho_k\gamma c$. 最后，我们注意到

$$P_1(G) = \gamma G + (1-\gamma)I$$ ■

正如我们在外推法中分析那样，我们可得到 $P_k(G)$谱半径的一个上界：

$$\rho(P_k(G)) = \max_{\lambda \in \Lambda(P_k(G))} |\lambda| = \max_{\lambda \in \Lambda(G)} |P_k(\lambda)|$$

$$\leqslant \max_{a \leqslant \lambda \leqslant b} |P_k(\lambda)| = 1/T_k(w(1))$$

在引理 2 中已作出上述要求．我们可以利用习题 4.6.36 的结果计算这个界并且得到这些公式：

$$\frac{1}{T_k(w(1))} = \frac{2}{b^n + b^{-n}} \quad b = t + \sqrt{t^2 - 1} \quad t = w(1)$$

可以证明切比雪夫加速比外推法快一个数量级．更多的细节见 Hageman and Young[1981]或 Kincaid and Young[1979]．

227

例 6 利用雅可比方法的切比雪夫加速求解以下问题：

$$\begin{bmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -4 \\ 0 \\ 4 \\ -4 \end{bmatrix}$$

解 首先我们用 $D^{-1}Ax = D^{-1}b$ 尺度化方程组，这里 $D = \text{diag}(A)$，得到雅可比迭代矩阵，其特征值位于区间$[-1/2, 1/2]$中．可应用切比雪夫加速法来加速这个基本方法的收敛性．一开始用初始向量 $u = (0, 0, 0, 0)^T$，利用类似于 Marc-32 的计算机，10 步迭代后收敛于近似解 $u = (-0.999\,996, -0.500\,002, 0.500\,002, -0.999\,996)^T$． ■

可写出**切比雪夫加速方法**的算法如下：

```
input u, a, b, M, δ
γ←2/(2−b−a)
α←[(1/2)(b−a)/(2−b−a)]²
output 0, u
call Extrap (γ, n, G, c, u, v)
output 1, v
ρ←1/(1−2α)
call Cheb(ρ, γ, n, G, c, u, v)
output 2, u
for k=3 to M step 2 do
    ρ←(1−ρα)
    call Cheb (ρ, γ, n, G, c, v, u)
    output k, v
    ρ←1/(1−ρα)
    call Cheb (ρ, γ, n, G, c, u, v)
    output k+1, u
    if ‖u−v‖∞<δ stop
end do
```

此处这个方法的第 1 步就是外推过程，每个相继的迭代包含两个加速步．执行基本的迭代步需要两个子程序或过程，即 Extrap 和 Cheb；它们在下面给出．注意通过适当的访问子程序

或过程 Cheb，我们即可得到所求向量的一个自动交换.

```
procedure Extrap (γ, n, G, c, u, v)
v←γc+(1−γ)u
v←γGu+v
return
procedure Cheb (ρ, γ, n, G, c, u, v)
u←ργc+ρ(1−γ)v+(1−ρ)u
u←ργGv+u
return
```

228

习题 4.6

1. 证明：若 A 对角占优且 Q 如在雅可比方法中那样选择，则

$$\rho(I-Q^{-1}A)<1$$

2. 证明：若 A 有性质(**单位行对角占优**)：

$$a_{ii}=1>\sum_{\substack{j=1\\j\neq i}}^{n}|a_{ij}|\qquad(1\leqslant i\leqslant n)$$

则理查森迭代成功.

3. (续)假设(**单位列对角占优**)：

$$a_{jj}=1>\sum_{\substack{i=1\\i\neq j}}^{n}|a_{ij}|\qquad(1\leqslant j\leqslant n)$$

重复上题.

4. (续)证明：若 A 有上面习题 2 中的性质，则下列迭代的极限是 $Ax=b$ 的解：

```
for k=1 to …
    for i=1 to n do
        x_i ← x_i + b_i − Σ_{j=1}^{n} a_ij x_j
    end do
end do
```

5. 设 $\|\cdot\|$ 是 $\mathbb{R}^n$ 上的一个范数，S 是 $n\times n$ 非奇异阵. 定义 $\|x\|'=\|Sx\|$，证明 $\|\cdot\|'$ 是一个范数.
6. (续)设 $\|\cdot\|$ 是一个从属矩阵范数，S 是一个非奇异阵. 定义 $\|A\|'=\|SAS^{-1}\|$，证明 $\|\cdot\|'$ 是一个从属矩阵范数.
7. 利用高斯-赛德尔方法中的 Q，证明：若 A 对角占优，则 $\|I-Q^{-1}A\|_\infty<1$.
8. 证明：$\rho(A)<1$ 当且仅当对每个 x，$\lim\limits_{k\to\infty}A^kx=0$.
9. 证明：若方程组 $Ax=b$ 中的第 i 个方程被 a_{ii} 除，然后应用理查森迭代，则这个结果与第一步应用雅可比迭代的结果相同.
10. 对谱半径函数 ρ，范数公理中哪几条是满足的？哪几条不满足？给出适当的证明和例子.
11. 对固定的 n，考虑上三角 $n\times n$ 矩阵组成的集合. 说明这个集合是向量空间. 并证明谱半径函数 ρ 是向量空间上的**拟范数**，因为除了 $A\neq 0$ 时 $\rho(A)$ 可能为 0 外，它满足所有的范数公理.

229

12. 说明为什么在定理 3 的证明中，我们不能设 $\epsilon\to 0$ 以及得到 A 相似于对角阵.
13. 设 A 可逆并设 f 是形如 $f(z)=\sum_{j=-m}^{m}c_jz^j$ 的函数. 证明：若 λ 是 A 的特征值，则 $f(\lambda)$ 是 $f(A)$ 的特征值.

14. 证明埃尔米特阵的特征值是实的. 提示：考虑$\langle x, Ax\rangle$和$\langle Ax, x\rangle$.

15. 设 A 对角占优，并设 Q 如高斯-赛德尔方法中的那样是 A 的下三角部分. 证明：$\rho(I-Q^{-1}A)$不大于下列比值中最大的值

$$r_i = \left\{\sum_{j=i+1}^{n} |a_{ij}|\right\} \Big/ \left\{|a_{ii}| - \sum_{j=1}^{i-1} |a_{ij}|\right\}$$

16. 证明：若 A 非奇异，则 AA^* 正定.

17. 证明：若 A 正定，则它的特征值是正的.

18. 证明：若 A 正定，则 A^2，A^3，…以及 A^{-1}，A^{-2}，…也正定.

19. 是否存在矩阵 A，对所有从属矩阵范数都满足 $\rho(A)<\|A\|$？

20. 证明：若 $\rho(A)<1$，则 $I-A$ 可逆且$(I-A)^{-1}=\sum_{k=0}^{\infty} A^k$.

21. 对一切 $n\times n$ 矩阵对 A 和 B，不等式 $\rho(AB)\leqslant\rho(A)\rho(B)$是否成立？当 A 和 B 是上三角阵时，你的回答是否相同？

22. 说明由(3)式给出的基本迭代与下列做法等价：给出 $x^{(k)}$，计算 $r^{(k)}=b-Ax^{(k)}$，在方程 $Qz^{(k)}=r^{(k)}$ 中解 $z^{(k)}$ 并定义 $x^{(k+1)}=x^{(k)}+z^{(k)}$.

23.(续)利用上题中的记号，说明

$$r^{(k+1)} = (I-AQ^{-1})r^{(k)}$$
$$z^{(k+1)} = (I-Q^{-1}A)z^{(k)}$$

24. n 阶埃尔米特阵是否能构成复域上的向量空间？

25. 说明：对非奇异矩阵 A 和 B，$\rho(AB)=\rho(BA)$.（如果可能，证明更强些的结果.）这个事实与上面的习题 4.6.23 有何关联？

26. 一个对角阵 D 对任意的 A，使得从 A 的正定性推出 DA 的正定性的充要条件是什么？

27. 证明：高斯-赛德尔方法是 SOR 方法的一个特殊情况.

28. 说明矩阵

$$\mathcal{R} = I - A$$
$$\mathcal{J} = I - D^{-1}A$$
$$\mathcal{G} = I - (D - C_L)^{-1}A$$
$$\mathcal{L}_\omega = I - \omega(D - \omega C_L)^{-1}A$$
$$\mathcal{U}_\omega = I - \omega(D - \omega C_U)^{-1}A$$
$$\mathcal{S}_\omega = I - \omega(2-\omega)(D - \omega C_U)^{-1}D(D - \omega C_L)^{-1}A$$

分别是理查森、雅可比、高斯-赛德尔、向前 SOR、向后 SOR 和 SSOR 方法的迭代矩阵. 然后说明本节中给出的分裂矩阵 Q 和迭代矩阵 G 是正确的.

230

29. 当

$$A = \begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix}$$

时，求高斯-赛德尔方法中迭代矩阵 $I-Q^{-1}A$ 的显式形式.

30. 对一切 $n\times n$ 非奇异矩阵 A，用 $x=0$ 作为初始向量，表示求解 $Ax=b$ 的高斯-赛德尔算法的一步的特性.

31. 举例说出一个非对角占优的矩阵 A，在高斯-赛德尔方法应用于 $Ax=b$ 时却收敛.

32. 若基本的方法是雅可比方法，问怎样简化切比雪夫加速方法?

33. 证明：若数 $\delta=\| I-Q^{-1}A\|$ 小于1，则

$$\| x^{(k)}-x\| \leqslant \frac{\delta}{1-\delta}\| x^{(k)}-x^{(k-1)}\|$$

34. 证明：

$$T_n(t)=\frac{1}{2}(b^n+b^{-n}) \qquad b=t+\sqrt{t^2-1}$$

35. 对 $a>1$ 的情况，证明定理9.

36. 利用定义证明正定矩阵是埃尔米特的，即对一切非零的 $x\in\mathbb{C}^n$，若 $x^*Ax>0$，则 A 是正定的. 因此，特别地，若利用这个定义，则实正定矩阵 A 一定对称. 然而，若使用另一个定义：对一切非零的 $x\in\mathbb{R}^n$，$x^TAx>0$，则实正定阵不一定对称.

37. 在相似等价关系下，所有相似于一个给定的 $n\times n$ 矩阵 A 的集合是一个等价类. 这些等价类是否为闭的? 因此，我们问由条件 $B^{(k)}\simeq A$ 和 $B^{(k)}\to B$ 是否推出 $B\simeq A$. 提示：见习题4.6.12).

38. 证明习题4.6.17的逆命题不成立.

39. 证明：若 A 非奇异且 $|\lambda|<\| A^{-1}\|^{-1}$，则 λ 不是 A 的特征值. 这里范数可以是任意从属矩阵范数.

计算机习题4.6

1. 对下列这些例子编写高斯-赛德尔方法的程序并进行测试.

a. $\begin{cases}3x+y+z=5\\x+3y-z=3\\3x+y-5z=-1\end{cases}$

b. $\begin{cases}3x+y+z=5\\3x+y-5z=-1\\x+3y-z=3\end{cases}$

231 当用不选主元的简单高斯消元法解这些方程组时，分析所发生的情况.

2. 应用高斯-赛德尔迭代于方程组，其中

$$A=\begin{bmatrix}0.963\,26 & 0.813\,21\\0.813\,21 & 0.686\,54\end{bmatrix} \qquad b=\begin{bmatrix}0.888\,24\\0.749\,88\end{bmatrix}$$

利用 $(0.331\,16,\ 0.700\,00)^T$ 作为初始向量，并说明发生的情况.

4.7 最速下降法和共轭梯度法

在本节中将讨论求解方程组

$$Ax=b$$

的某些特殊方法. 此时，A 是一个实 $n\times n$ **对称和正定阵**. 这些假设意指

$$A^T=A$$

且

$$x^TAx>0, x\neq 0$$

我们始终利用实向量 x 和 y 的内积：

$$\langle x,y\rangle = x^{\mathrm{T}}y = \sum_{i=1}^{n} x_i y_i$$

一些直接得到的性质有：

1. $\langle x, y\rangle = \langle y, x\rangle$.
2. 对任意常数 α，$\langle \alpha x, y\rangle = \alpha\langle x, y\rangle$.
3. $\langle x+y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$.
4. $\langle x, Ay\rangle = \langle A^{\mathrm{T}}x, y\rangle$.

由性质 1 知，性质 2 和性质 3 中的变量次序可以相反.

我们开始建立涉及 A 和 b 两个数值问题的等价性.

引理 1（二次型引理） 若 A 对称正定，则求解 $Ax=b$ 的问题等价于极小化二次型问题

$$q(x) = \langle x,Ax\rangle - 2\langle x,b\rangle$$

证明 首先，让我们查明函数 q 沿**一维射线**如何变化. 这里我们考虑 $x+tv$，其中 x 和 v 是向量而 t 是纯量. 图 4-2 说明为什么可以认为 $x+tv$ 是一条一维射线. 232

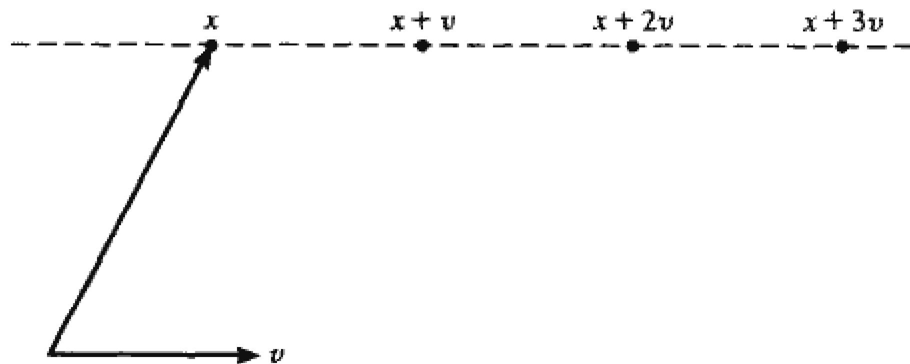


图 4-2 一维射线的例子

对纯量 t 直接展开计算，因为 $A^{\mathrm{T}}=A$，我们有

$$\begin{aligned} q(x+tv) &= \langle x+tv, A(x+tv)\rangle - 2\langle x+tv, b\rangle \\ &= \langle x,Ax\rangle + t\langle x,Av\rangle + t\langle v,Ax\rangle + t^2\langle v,Av\rangle - 2\langle x,b\rangle - 2t\langle v,b\rangle \\ &= q(x) + 2t\langle v,Ax\rangle - 2t\langle v,b\rangle + t^2\langle v,Av\rangle \\ &= q(x) + 2t\langle v,Ax-b\rangle + t^2\langle v,Av\rangle \end{aligned} \tag{1}$$

注意(1)式中 t^2 的系数是正的. 因此，射线上的二次函数有极小而不是极大值. 由(1)式，我们计算关于 t 的导数

$$\frac{\mathrm{d}}{\mathrm{d}t}q(x+tv) = 2\langle v,Ax-b\rangle + 2t\langle v,Av\rangle \tag{2}$$

当(2)中的导数为 0 时，q 沿射线出现极小. 所以产生极小值的 t 值为

$$\hat{t} = \langle v,b-Ax\rangle / \langle v,Av\rangle \tag{3}$$

利用这个值，我们在射线上计算 q 的极小值：

$$\begin{aligned} q(x+\hat{t}v) &= q(x) + \hat{t}[2\langle v,Ax-b\rangle + \hat{t}\langle v,Av\rangle] \\ &= q(x) + \hat{t}[2\langle v,Ax-b\rangle + \langle v,b-Ax\rangle] \\ &= q(x) - \hat{t}\langle v,b-Ax\rangle \end{aligned}$$

$$= q(x) - \langle v, b - Ax \rangle^2 / \langle v, Av \rangle \tag{4}$$

计算表明，除非 v **正交**于残差，即$\langle v, b-Ax\rangle=0$，否则经过 x 到 $x+\hat{t}v$，q 的值总是出现减少．如果 x 不是方程组 $Ax=b$ 的解，那么存在许多向量 v 满足$\langle v, b-Ax\rangle\neq0$．因此，若 $Ax\neq b$，则 x 不极小化 q．另一方面，若 $Ax=b$，则不存在从 x 出发的这种射线，在其上 q 的值比 $q(x)$更小．所以，这样的 x 是 q 的极小点． ■

233

上面的证明让我们联想到一个求解 $Ax=b$ 的迭代法．我们通过沿着一系列射线极小化 q 进行下去．在这样的一个算法中的第 k 步，可得到 $x^{(0)}$，$x^{(1)}$，$x^{(2)}$，…，$x^{(k)}$．然后，利用某种规则，选择一个适当的搜索方向 $v^{(k)}$．则在我们的序列中的下一点就是

$$x^{(k+1)} = x^{(k)} + t_k v^{(k)}$$

其中

$$t_k = \frac{\langle v^{(k)}, b - Ax^{(k)} \rangle}{\langle v^{(k)}, Av^{(k)} \rangle}$$

对纯量值 t_k 和向量 $v^{(k)}$的特定值，许多迭代法具有一般形式

$$x^{(k+1)} = x^{(k)} + t_k v^{(k)}$$

若$\| v^{(k)} \|=1$，则 t_k 测量我们从 $x^{(k)}$移动到 $x^{(k+1)}$的距离．

4.7.1 最速下降法

最速下降法是刚才所述类型的一个算法．它约定 $v^{(k)}$是 q 在 $x^{(k)}$的负梯度．这产生残差 $r^{(k)}=b-Ax^{(k)}$方向的负梯度点．（见习题 4.7.1.）有关最速下降的形式描述如下：

```
input x^(0), A, b, M
output 0, x^(0)
for k=0 to M-1 do
    v^(k) ← b - Ax^(k)
    t_k ← ⟨v^(k), v^(k)⟩/⟨v^(k), Av^(k)⟩
    x^(k+1) ← x^(k) + t_k v^(k)
output k+1, x^(k+1)
end do
```

在这个算法的实际程序设计中，逐次产生的向量 $x^{(0)}$，$x^{(1)}$，…不需要存储；而当前的 x 向量可以覆盖．同样的论点对方向向量 $v^{(0)}$，$v^{(1)}$，…也成立．因而，我们又可以写成

```
input x, A, b, M
output 0, x
for k=1 to M do
    v ← b - Ax
    t ← ⟨v, v⟩/⟨v, Av⟩
    x ← x + tv
    output k, x
end do
```

234

因为太慢了，所以最速下降法在这个问题中极少使用．图 4-3 中描述了对二维问题它是如何进行的，其中我们显示了二次型 q 的等高线（或**水平线**）．

图 4-3 最速下降法的几何解释

4.7.2 共轭方向

一族称为**共轭方向法**的方法都采用沿一系列射线极小化二次函数的基本策略. 通常这些搜索方向是在求解过程中一个一个确定的. 然而，我们将首先讨论在开始时它们就被指定下来的情况.

假设 A 是 $n\times n$ 对称正定阵，假如提供一组向量 $\{u^{(1)}, u^{(2)}, \cdots, u^{(n)}\}$ 且有性质

$$\langle u^{(i)}, Au^{(j)}\rangle = \delta_{ij} \qquad (1\leqslant i,j\leqslant n)$$

我们称这个性质为 **A 标准正交性**，并且显然它是通常正交性的推广. 我们将看到若这些向量在二次函数 q 一步步极小化中用作搜索方向，则由第 n 步得到解. 在给出结果之前，我们注意到 A 标准正交性条件可以用矩阵等式

$$U^{\mathrm{T}}AU = I$$

来表示，其中 U 是列为 $u^{(1)}$, $u^{(2)}$, $\cdots$, $u^{(n)}$ 的 $n\times n$ 矩阵. 显然由此可得 A 和 U 是非奇异的，且列 $u^{(1)}$, $u^{(2)}$, $\cdots$, $u^{(n)}$ 构成$\mathbb{R}^n$的一个基底.

定理 1（A-标准正交系定理） 设 $\{u^{(1)}, u^{(2)}, \cdots, u^{(n)}\}$ 是一个 A 标准正交系. 定义

$$x^{(i)} = x^{(i-1)} + \langle b - Ax^{(i-1)}, u^{(i)}\rangle u^{(i)} \qquad (1\leqslant i\leqslant n)$$

其中 $x^{(0)}$ 是$\mathbb{R}^n$中的任意点，则 $Ax^{(n)}=b$.

证明 定义 $t_i=\langle b-Ax^{(i-1)}, u^{(i)}\rangle$，使递归公式简化为

$$x^{(i)} = x^{(i-1)} + t_i u^{(i)}$$

[235]

由此并借助于关系 $\langle Au^{(j)}, u^{(i)}\rangle=\delta_{ij}$，我们导出下列关系：

$$Ax^{(i)} = Ax^{(i-1)} + t_i Au^{(i)}$$

$$Ax^{(n)} = Ax^{(0)} + t_1 Au^{(1)} + \cdots + t_n Au^{(n)}$$

$$\langle Ax^{(n)} - b, u^{(i)}\rangle = \langle Ax^{(0)} - b, u^{(i)}\rangle + t_i$$

下面，我们指出最后等式的右边为 0：

$$\begin{aligned} t_i &= \langle b - Ax^{(i-1)}, u^{(i)}\rangle \\ &= \langle b - Ax^{(0)}, u^{(i)}\rangle + \langle Ax^{(0)} - Ax^{(1)}, u^{(i)}\rangle + \cdots + \langle Ax^{(i-2)} - Ax^{(i-1)}, u^{(i)}\rangle \\ &= \langle b - Ax^{(0)}, u^{(i)}\rangle + \langle -t_1 Au^{(1)}, u^{(i)}\rangle + \cdots + \langle -t_{i-1} Au^{(i-1)}, u^{(i)}\rangle \\ &= \langle b - Ax^{(0)}, u^{(i)}\rangle \end{aligned}$$

我们的分析表明 $Ax^{(n)}-b$（在通常意义下）正交于 $u^{(1)}$, $u^{(2)}$, $\cdots$, $u^{(n)}$，所以 $Ax^{(n)}-b$ 必为 0. ■

现在我们描述这个方法的具体实现. 设 A 是 $n\times n$ 对称正定矩阵. 等式

$$\langle x,y\rangle_A = \langle x,Ay\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_iy_j$$

定义一个内积，这是很容易验证的．存在一个与之相应的**二次范数**

$$\|x\|_A^2 = \langle x,x\rangle_A$$

如果格拉姆-施密特过程以这个新内积用于标准单位向量组$\{e^{(1)}, e^{(2)}, \cdots, e^{(n)}\}$，结果就是一个 A 标准正交系$\{u^{(1)}, u^{(2)}, \cdots, u^{(n)}\}$．描述格拉姆-施密特过程(更详细的讨论在 5.3 节)如下：

$$u^{(i)} = \|v^{(i)}\|_A^{-1}v^{(i)}, \text{这里 } v^{(i)} = e^{(i)} - \sum_{j<i}\langle e^{(i)},u^{(j)}\rangle_A u^{(j)}$$

因为向量 $e^{(i)}$ 的特性，这些公式化成

$$u^{(i)} = \|v^{(i)}\|_A^{-1}v^{(i)}, \text{这里 } v^{(i)} = e^{(i)} - \sum_{j<i}(Au^{(j)})_i u^{(j)}$$

例如，$u^{(1)}=(1/\sqrt{a_{11}})e^{(1)}$．这个公式指出每个 $u^{(i)}$ 是单位向量 $e^{(1)}, e^{(2)}, \cdots, e^{(i)}$ 的线性组合．所以，列向量为 $u^{(1)}, u^{(2)}, \cdots, u^{(n)}$ 的矩阵 U 是上三角阵．上面提及的等式 $U^TAU=I$ 给出 $A=(U^T)^{-1}U^{-1}$，这是一个 LU 分解．于是 $Ax=b$ 的解可用不选主元的高斯消元法的一种变形算出来．

236

在数值计算工作中，更为方便的是使用 A 正交系而不是 A 标准正交系．如果对 $i\neq j$，有 $\langle v^{(i)}, Av^{(j)}\rangle=0$，那么称向量组 $v^{(1)}, v^{(2)}, \cdots$为 **A 正交的**．我们用规范化过程

$$u^{(i)} = \|v^{(i)}\|_A^{-1}v^{(i)}$$

把一个 A 正交系转换为一个 A 标准正交系．若每个 $v^{(i)}$ 是非零向量且 A 是正定阵，则 $u^{(i)}$ 将构成 A 标准正交系．下面的定理类似于定理 1.

定理 2(A 正交系定理)　*设$\{v^{(1)}, v^{(2)}, \cdots, v^{(n)}\}$是对称正定 $n\times n$ 矩阵 A 的非零向量 A 正交系．定义*

$$x^{(i)} = x^{(i-1)} + \frac{\langle b-Ax^{(i-1)}, v^{(i)}\rangle}{\langle v^{(i)}, Av^{(i)}\rangle}v^{(i)} \qquad (1\leqslant i\leqslant n)$$

其中 $x^{(0)}$ 是任意的，则 $Ax^{(n)}=b$.

4.7.3 共轭梯度法

Hestenes and Stiefel[1952]的**共轭梯度法**是一种特殊类型的共轭方向法．它应用于方程组 $Ax=b$，其中 A 是对称正定阵．在共轭梯度法中，搜索方向 $v^{(i)}$ 参照定理 2 一个接一个在迭代过程中选取并构成 A 正交系．然而，特性是残差 $r^{(i)}=b-Ax^{(i)}$ 构成了通常意义下的正交系；即$\langle r^{(i)}, r^{(j)}\rangle=0$，$i\neq j$.

当 A 是一个非常大的稀疏阵时，共轭梯度法比简单的高斯消元法更值得推荐．理论上，共轭梯度算法至多需要 n 步就能得到方程组 $Ax=b$ 的解．然而，实际上算法是作为一个迭代法来使用，产生收敛于解的一个向量序列．在病态问题中，舍入误差通常妨碍算法在第 n 步提供一个充分精确的解．当共轭梯度法由 Hestenes and Stiefel[1952]引进时，一开始，人们对它非常兴奋．然而，当发现它的有限终止性质实际上达不到时，人们的兴趣很快地减少了．因为作为一个**直接法**，这是不符合要求的．20 年以后，当它被看作**迭代法**时，人们对这个方法又

重新产生了兴趣. 在 n 步之后，得不到完全精确的解是迭代法所预期的情况. 实际上，对极大型方程组，我们希望在比 n 稍微多一些步数能给出一个满意的解答. 对良态问题，为了使共轭梯度法有令人满意的收敛性，所需要的迭代步数可能稍为比方程组的阶要小些. Golub and O'Leary[1989]已撰写了这个方法的历史.

形式共轭梯度算法如下.（这个伪代码不是为计算机执行所设计的. 为那个目的相配的算法将在后面给出.）

237

```
input x^(0), M, A, b, ε
r^(0) ← b − Ax^(0)
v^(0) ← r^(0)
output 0, x^(0), r^(0)
for k=0 to M−1 do
    if v^(k) = 0 then stop
    t_k ← ⟨r^(k), r^(k)⟩/⟨v^(k), Av^(k)⟩
    x^(k+1) ← x^(k) + t_k v^(k)
    r^(k+1) ← r^(k) − t_k A v^(k)
    if ‖r^(k+1)‖_2^2 < ε then stop
    s_k ← ⟨r^(k+1), r^(k+1)⟩/⟨r^(k), r^(k)⟩
    v^(k+1) ← r^(k+1) + s_k v^(k)
    output k+1, x^(k+1), r^(k+1)
end do
```

在算法中，若 $r^{(k)}=0$，则 $x^{(k)}$(理论上)是线性方程组 $Ax=b$ 的解，算法的计算机实现需要 4 个向量 $x^{(k)}$，$r^{(k)}$，$v^{(k)}$，$Av^{(k)}$. 为计算积 $Av^{(k)}$，也必须做好准备.（可能需要或不需要存储整个矩阵 A.）每次迭代工作量是适度的，等于单个矩阵-向量积 $Av^{(k)}$ 和(第 1 次迭代后)两个内积 $\langle v^{(k)}, Av^{(k)}\rangle$ 和 $\langle r^{(k+1)}, r^{(k+1)}\rangle$. 注意，停止检验涉及 $\|r^{(k+1)}\|_2^2=\langle r^{(k+1)}, r^{(k+1)}\rangle$，它很容易计算，因为已经得到 $r^{(k+1)}$. 所以**共轭梯度法**的计算机代码应该基于下列算法：

```
input x, A, b, M, ε, δ
r ← b − Ax
v ← r
c ← ⟨r, r⟩
for k=1 to M do
    if ⟨v, v⟩^(1/2) < δ then exit loop
    z ← Av
    t ← c/⟨v, z⟩
    x ← x + tv
    r ← r − tz
    d ← ⟨r, r⟩
    if d < ε then exit loop
    v ← r + (d/c)v
    c ← d
    output k, x, r
end do
```

定理 3(共轭梯度算法定理) 在共轭梯度算法中，对任意的整数 $m<n$. 若 $v^{(0)}$, $v^{(1)}$, …, $v^{(m)}$ 全部是非零向量，则 $r^{(i)}=b-Ax^{(i)}$, $0\leqslant i\leqslant m$, 且 $\{r^{(0)}, r^{(1)}, \cdots, r^{(m)}\}$ 是一个非零向量的正交集合.

238

证明 我们将在给定的假设下证明下面每一个式子都成立：

1. $\langle r^{(m)}, v^{(i)}\rangle=0(0\leqslant i<m)$.
2. $\langle r^{(i)}, r^{(i)}\rangle=\langle r^{(i)}, v^{(i)}\rangle(0\leqslant i\leqslant m)$.
3. $\langle v^{(m)}, Av^{(i)}\rangle=0(0\leqslant i<m)$.
4. $r^{(i)}=b-Ax^{(i)}(0\leqslant i\leqslant m)$.
5. $\langle r^{(m)}, r^{(i)}\rangle=0(0\leqslant i<m)$.
6. $r^{(i)}\neq 0(0\leqslant i\leqslant m)$.

对 m 用归纳法来证明. $m=0$ 时，假定 $v^{(0)}\neq 0$，并且我们必须证明 1～6. 注意此时第 1，3，5 是无关的，至于第 2，4，6，它们从算法的定义直接可得，因为

$$r^{(0)}=b-Ax^{(0)}=v^{(0)}\neq 0$$

现在假设定理已经对某个 m 成立. 在此基础上，我们将证明它对 $m+1$ 也成立. 此时，我们假设 $v^{(0)}, v^{(1)}, \cdots, v^{(m+1)}$ 全不为 0. 由归纳假定，第 1～6 成立. 我们需要证明

1′. $\langle r^{(m+1)}, v^{(i)}\rangle=0(0\leqslant i\leqslant m)$.

2′. $\langle r^{(m+1)}, r^{(m+1)}\rangle=\langle r^{(m+1)}, v^{(m+1)}\rangle$.

3′. $\langle v^{(m+1)}, Av^{(i)}\rangle=0(0\leqslant i\leqslant m)$.

4′. $r^{(m+1)}=b-Ax^{(m+1)}$.

5′. $\langle r^{(m+1)}, r^{(i)}\rangle=0(0\leqslant i\leqslant m)$.

6′. $r^{(m+1)}\neq 0$.

为证明第 1′，首先设 $i=m$. 利用第 2 我们有

$$\begin{aligned}\langle r^{(m+1)},v^{(m)}\rangle&=\langle r^{(m)}-t_mAv^{(m)},v^{(m)}\rangle=\langle r^{(m)},v^{(m)}\rangle-t_m\langle v^{(m)},Av^{(m)}\rangle\\&=\langle r^{(m)},v^{(m)}\rangle-\langle r^{(m)},r^{(m)}\rangle=0\end{aligned}$$

若 $0\leqslant i<m$，则由第 1 和第 3 可得

$$\langle r^{(m+1)},v^{(i)}\rangle=\langle r^{(m)},v^{(i)}\rangle-t_m\langle v^{(m)},Av^{(i)}\rangle=0$$

对第 2′的证明，我们可利用第 1′得出

$$\langle r^{(m+1)},v^{(m+1)}\rangle=\langle r^{(m+1)},r^{(m+1)}+s_mv^{(m)}\rangle=\langle r^{(m+1)},r^{(m+1)}\rangle$$

在第 3′证明中，当 s_{-1} 和 $v^{(-1)}$ 出现时，我们令 $s_{-1}=0$ 和 $v^{(-1)}=0$. 若 $0\leqslant i\leqslant m$，则

239

$$\begin{aligned}\langle v^{(m+1)},Av^{(i)}\rangle&=\langle r^{(m+1)}+s_mv^{(m)},Av^{(i)}\rangle\\&=\langle r^{(m+1)},Av^{(i)}\rangle+s_m\langle v^{(m)},Av^{(i)}\rangle\\&=t_i^{-1}\langle r^{(m+1)},r^{(i)}-r^{(i+1)}\rangle+s_m\langle v^{(m)},Av^{(i)}\rangle\\&=t_i^{-1}\langle r^{(m+1)},v^{(i)}-s_{i-1}v^{(i-1)}-v^{(i+1)}+s_iv^{(i)}\rangle+s_m\langle v^{(m)},Av^{(i)}\rangle\\&=t_i^{-1}[\langle r^{(m+1)},v^{(i)}\rangle-s_{i-1}\langle r^{(m+1)},v^{(i-1)}\rangle-\langle r^{(m+1)},v^{(i+1)}\rangle\\&\quad+s_i\langle r^{(m+1)},v^{(i)}\rangle]+s_m\langle v^{(m)},Av^{(i)}\rangle\end{aligned}$$

若 $i<m$，则由第 1′知，$r^{(m+1)}$ 正交于 $v^{(i)}$, $v^{(i-1)}$, $v^{(i+1)}$. 由第 3 知，$\langle v^{(m)}, Av^{(i)}\rangle=0$. 因此，在

这种情况中，$\langle v^{(m+1)}, Av^{(i)}\rangle=0$. 而 $i=m$ 情况是较特别的. 由前面的等式，我们有

$$\langle v^{(m+1)},Av^{(m)}\rangle = t_m^{-1}\langle r^{(m+1)},v^{(m)} - s_{m-1}v^{(m-1)} - v^{(m+1)} + s_m v^{(m)}\rangle$$
$$+ s_m\langle v^{(m)},Av^{(m)}\rangle$$

由第 1′，$r^{(m+1)}$ 正交于 $v^{(m)}$ 和 $v^{(m-1)}$. 因此，

$$\langle v^{(m+1)},Av^{(m)}\rangle = -t_m^{-1}\langle r^{(m+1)},v^{(m+1)}\rangle + s_m\langle r^{(m)},Av^{(m)}\rangle$$
$$= -\frac{\langle v^{(m)},Av^{(m)}\rangle}{\langle r^{(m)},r^{(m)}\rangle}\langle r^{(m+1)},v^{(m+1)}\rangle + \frac{\langle r^{(m+1)},r^{(m+1)}\rangle}{\langle r^{(m)},r^{(m)}\rangle}\langle v^{(m)},Av^{(m)}\rangle$$

利用第 2′，我们看出这个表达式为 0.

对第 4′的证明，我们记

$$b - Ax^{(m+1)} = b - A(x^{(m)} + t_m v^{(m)}) = b - Ax^{(m)} - t_m Av^{(m)}$$
$$= r^{(m)} - (r^{(m)} - r^{(m+1)}) = r^{(m+1)}$$

对第 5′的证明，设 $0\leqslant i\leqslant m$ 并令 $s_{-1}=0$ 和 $v^{(-1)}=0$. 则从第 1′我们有

$$\langle r^{(m+1)},r^{(i)}\rangle = \langle r^{(m+1)},v^{(i)} - s_{i-1}v^{(i-1)}\rangle$$
$$= \langle r^{(m+1)},v^{(i)}\rangle - s_{i-1}\langle r^{(m+1)},v^{(i-1)}\rangle = 0$$

对第 6′，利用第 3′和 A 的正定性证明如下：

$$0 < \langle v^{(m+1)},Av^{(m+1)}\rangle = \langle r^{(m+1)} + s_m v^{(m)},Av^{(m+1)}\rangle$$
$$= \langle r^{(m+1)},Av^{(m+1)}\rangle + s_m\langle v^{(m)},Av^{(m+1)}\rangle$$
$$= \langle r^{(m+1)},Av^{(m+1)}\rangle$$

因此，$r^{(m+1)}\neq 0$. ■

本定理及其证明取自 Stoer and Bulirsch[1980].

4.7.4 预处理的共轭梯度法

我们要用共轭梯度法的一个变形求解方程组 [240]

$$Ax = b$$

其中 A 是对称正定阵. **预处理**这个方程组，并得到一个比原方程组有更好条件的新方程组将是十分有益的. 我们意指对某个非奇异阵 S，预处理的方程组

$$\hat{A}\hat{x} = \hat{b}$$

其中

$$\begin{cases}\hat{A} = S^{\mathrm{T}}AS\\ \hat{x} = S^{-1}x\\ \hat{b} = S^{\mathrm{T}}b\end{cases}$$

使得 $\kappa(\hat{A})<\kappa(A)$. 作为一个副产品，用于求解预处理方程组的迭代法比它用于求解原方程组可能要收敛得快. 那么与其取任意的 S，不如假定对称正定分裂矩阵 Q 可分解成

$$Q^{-1} = SS^{\mathrm{T}}$$

这样做的原因不久就会清楚.

对预处理的方程组可写出如下形式的共轭梯度算法：

$\hat{r}^{(0)}=\hat{b}-\hat{A}\hat{x}^{(0)}$
$\hat{v}^{(0)}=\hat{r}^{(0)}$
for $k=0$ **to** M **do**
  $\hat{t}_k=\langle\hat{r}^{(k)},\ \hat{r}^{(k)}\rangle/\langle\hat{v}^{(k)},\ \hat{A}\hat{v}^{(k)}\rangle$
  $\hat{x}^{(k+1)}=\hat{x}^{(k)}+\hat{t}_k\hat{v}^{(k)}$
  $\hat{r}^{(k+1)}=\hat{r}^{(k)}-\hat{t}_k\hat{A}\hat{v}^{(k)}$
  $\hat{s}_k=\langle\hat{r}^{(k+1)},\ \hat{r}^{(k+1)}\rangle/\langle\hat{r}^{(k)},\ \hat{r}^{(k)}\rangle$
  $\hat{v}^{(k+1)}=\hat{r}^{(k+1)}+\hat{s}_k\hat{v}^{(k)}$
end do

在预处理原方程组过程中，矩阵 A 中的稀疏性在形成 $\hat{A}$ 时可能被破坏. 故与其显式地构成预处理方程组，不如我们利用原方程组并在算法中隐式地做预处理.

记

$$
\begin{aligned}
\hat{x}^{(k)} &= S^{-1}x^{(k)} \\
\hat{v}^{(k)} &= S^{-1}v^{(k)} \\
\hat{r}^{(k)} &= \hat{b}-\hat{A}\hat{x}^{(k)} = S^{\mathrm{T}}b-(S^{\mathrm{T}}AS)(S^{-1}x^{(0)}) = S^{\mathrm{T}}r^{(k)} \\
\tilde{r}^{(k)} &= Q^{-1}r^{(k)}
\end{aligned}
$$

[241] 于是，得到

$$
\begin{aligned}
\hat{t}_k &= \langle\hat{r}^{(k)},\hat{r}^{(k)}\rangle/\langle\hat{v}^{(k)},\hat{A}\hat{v}^{(k)}\rangle \\
&= \langle S^{\mathrm{T}}r^{(k)},S^{\mathrm{T}}r^{(k)}\rangle/\langle S^{-1}v^{(k)},(S^{\mathrm{T}}AS)(S^{-1}v^{(k)})\rangle \\
&= \langle Q^{-1}r^{(k)},r^{(k)}\rangle/\langle v^{(k)},Av^{(k)}\rangle \\
&= \langle\tilde{r}^{(k)},r^{(k)}\rangle/\langle v^{(k)},Av^{(k)}\rangle
\end{aligned}
$$

从而，我们有

$$
\begin{aligned}
\hat{x}^{(k+1)} &= \hat{x}^{(k)}+\hat{t}_k\hat{v}^{(k)} \\
S^{-1}x^{(k+1)} &= S^{-1}x^{(k)}+\hat{t}_kS^{-1}v^{(k)}
\end{aligned}
$$

用 S 相乘，我们得到

$$
x^{(k+1)} = x^{(k)}+\hat{t}_kv^{(k)}
$$

类似地，我们有

$$
\begin{aligned}
\hat{r}^{(k+1)} &= \hat{r}^{(k)}-\hat{t}_k\hat{A}\hat{v}^{(k)} \\
S^{\mathrm{T}}r^{(k+1)} &= S^{\mathrm{T}}r^{(k)}-\hat{t}_k(S^{\mathrm{T}}AS)(S^{-1}v^{(k)})
\end{aligned}
$$

用 $S^{-\mathrm{T}}$ 相乘，得到

$$
r^{(k+1)} = r^{(k)}-\hat{t}_kAv^{(k)}
$$

现在，我们有

$$
\begin{aligned}
\hat{s}_k &= \langle\hat{r}^{(k+1)},\hat{r}^{(k+1)}\rangle/\langle\hat{r}^{(k)},\hat{r}^{(k)}\rangle \\
&= \langle S^{\mathrm{T}}r^{(k+1)},S^{\mathrm{T}}r^{(k+1)}\rangle/\langle S^{\mathrm{T}}r^{(k)},S^{\mathrm{T}}r^{(k)}\rangle \\
&= \langle Q^{-1}r^{(k+1)},r^{(k+1)}\rangle/\langle Q^{-1}r^{(k)},r^{(k)}\rangle \\
&= \langle\tilde{r}^{(k+1)},r^{(k+1)}\rangle/\langle\tilde{r}^{(k)},r^{(k)}\rangle
\end{aligned}
$$

此外，还有

$$
\hat{v}^{(k+1)} = \hat{r}^{(k+1)}+\hat{s}_k\hat{v}^{(k)}
$$

$$S^{-1}v^{(k+1)} = S^{\mathrm{T}}r^{(k+1)} + \hat{s}_k S^{-1}v^{(k)}$$

用 S 相乘，我们得到

$$\begin{aligned} v^{(k+1)} &= Q^{-1}r^{(k+1)} + \hat{s}_k v^{(k)} \\ &= \tilde{r}^{(k+1)} + \hat{s}_k v^{(k)} \end{aligned}$$

现在我们主要根据原来的方程组编写**预处理共轭梯度算法**. 下面给出的伪代码不适合作为计算机程序. 另一个更为有效的形式在本节后面给出.

242

```
input x^(0), A, b, M, Q
r^(0) ← b − Ax^(0)
solve Q r̃^(0) = r^(0) for r̃^(0)
v^(0) ← r̃^(0)
output 0, x^(0)
for k=0 to M−1 do
    if v^(k) = 0 then exit loop
    t̂_k ← ⟨r̃^(k), r^(k)⟩/⟨v^(k), Av^(k)⟩
    x^(k+1) ← x^(k) + t̂_k v^(k)
    r^(k+1) ← r^(k) − t̂_k A v^(k)
    solve Q r̃^(k+1) = r^(k+1) for r̃^(k+1)
    if ⟨r̃^(k+1), r^(k+1)⟩ < ε then
        if ⟨r^(k+1), r^(k+1)⟩ < ε then exit loop
    end if
    ŝ_k ← ⟨r̃^(k+1), r^(k+1)⟩/⟨r̃^(k), r^(k)⟩
    v^(k+1) ← r̃^(k+1) + ŝ_k v^(k)
    output k+1, x^(k+1), r^(k+1)
end do
```

当 $Q^{-1}=I$ 时，因为 $\tilde{r}^{(k)}=r^{(k)}$，$\hat{t}_k=t_k$，$\hat{s}_k=s_k$，所以上面的预处理共轭梯度算法化为正则的共轭梯度算法.

若 $Q=A$，$S=A^{-1/2}$，则预处理方程组化为 $\hat{x}=\hat{b}$，将被平凡地解出. 遗憾的是，这个理想的条件方程组($\kappa(\hat{A})=1$)不具有计算价值，因为确定 $\hat{b}=S^{\mathrm{T}}b$ 将像解原方程组一样困难.

因为在预处理共轭梯度算法的每一步我们都必须解一个形如 $Qx=y$ 的方程组，所以要选择 Q 使得这个方程组容易求解. 若 Q 是对角阵，则满足这个条件，而更复杂的预处理可能导致更快的收敛性. 当 Q^{-1}变成 A 的较好的近似时，预处理方程组就具有更好的条件，并且只要少数几步就会出现迭代过程的收敛性. 另一方面，求解 $Qx=y$ 变得更为复杂；这说明了两种不同迭代法之间典型的利弊关系：一种是迭代步数较少但每次迭代计算量大，而另一种是迭代步数多但每次迭代计算量小.

怎样考虑收敛性的检验？在算法中 $\|r^{(k+1)}\|_2^2$ 不能利用，因为需要一个额外的计算. 可利用的是$\langle\tilde{r}^{(k+1)}, r^{(k+1)}\rangle$. 利用它可以使迭代过程在其达到指定的精确度之前或之后停止. 因此，一旦前面的检验指出收敛时，作为检验应该做额外的计算 $\|r^{(k+1)}\|_2^2$.

计算机代码可以基于下列**预处理共轭梯度算法**：

```
input x, A, b, M, Q, δ, ε
r ← b − Ax
```

```
solve Qz=r for z
v←z
c←⟨z, r⟩
for k=1 to M do
    if⟨v, v⟩^{1/2}<δ then exit loop
    z←Av
    t←c/⟨v, z⟩
    x←x+tv
    r←r−tz
    solve Qz=r for z
    d←⟨z, r⟩
    if d<ε then
        if ⟨r, r⟩<ε then exit loop
    end if
    v←z+(d/c)v
    c←d
    output k, x, r
end do
```

243

预处理共轭梯度法是一个活跃的研究领域. Q 有若干种选择，如对称逐次超松弛矩阵. 可以利用包含共轭梯度程序和其他迭代法的计算机软件包. 这些软件包的例子是 Kincaid, Oppe, and Young[1989]的 ITPACKV 2D、Oppe, Joubert, and Kincaid[1988]的 NSPCG、PCGPAK2[1990]、Joubert et al. [1995]的 PCG. 我们已经仿照 Ortega[1988]的表示在这里给出预处理共轭梯度法的表示. 更多的细节可以在那里和许多其他著作，如 Golub and Van Loan[1989]的书中找到.

习题 4.7

1. 证明：若 A 对称，则函数 $q(x)=\langle x, Ax\rangle-2\langle x, b\rangle$ 在 x 的梯度是 $2(Ax-b)$. 记函数 g: $\mathbb{R}^n\to\mathbb{R}$ 的梯度是分量为 $\partial g/\partial x_i$, $i=1, 2, \cdots, n$, 的向量.
2. 证明：$q(x)$的极小值是 $-\langle b, A^{-1}b\rangle$.
3. 证明：在最速下降法中 $v^{(k)}\perp v^{(k+1)}$.
4. 设 A 正定，且 b 是一个固定的向量. 对任意的 x，**残差向量**是 $r=b-Ax$，**误差向量**是 $e=A^{-1}b-x$. 说明：除非 $Ax=b$，否则误差向量与残差向量的内积是正的.
5. 设 A 对称，且 $Ax=b$ 而 y 是任意向量. 利用课本中定义的 q 证明
$$\langle(x-y),A(x-y)\rangle=\langle b,A^{-1}b\rangle+q(y)$$
这表明极小化 $q(y)$等价于极小化$\langle(x-y), A(x-y)\rangle$.
6. 证明：若 $\hat{t}$ 由(3)式定义且 $y=x+\hat{t}v$，则 $v\perp(b-Ay)$；即$\langle v, b-Ay\rangle=0$.
7. 说明在最速下降法中
$$q(x^{(k+1)})=q(x^{(k)})-\|r^{(k)}\|^4/\langle r^{(k)},Ar^{(k)}\rangle$$
其中 $r^{(k)}=b-Ax^{(k)}$.

244

8. 若$\{u^{(1)}, u^{(2)}, \cdots, u^{(n)}\}$是一组通常意义下的标准正交向量，如果把这些向量用于极小化 q 的搜索方向，那么 n 步以后是否会得到解？
9. 设 A 是 $n\times n$ 矩阵，不假设 A 是对称或正定阵. 假设存在 A 标准正交系$\{u^{(1)}, u^{(2)}, \cdots, u^{(n)}\}$，证明 A 是

对称正定阵.

10. 证明：在共轭梯度法中

$$t_k = \langle r^{(k)}, v^{(k)} \rangle / \langle v^{(k)}, Av^{(k)} \rangle$$
$$s_k = -\langle r^{(k+1)}, Av^{(k)} \rangle / \langle v^{(k)}, Av^{(k)} \rangle$$

11. 在共轭梯度法中，证明：若 $v^{(k)}=0$，则 $Ax^{(k)}=b$.

12. 对 t_k 的什么值，最速下降法化为理查森方法？为使最速下降法等价于雅可比法 t_k 和 A 必须满足什么条件？

13. 考虑线性方程组

$$\begin{bmatrix} 2 & 0 & -1 \\ -2 & -10 & 0 \\ -1 & -1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -12 \\ 2 \end{bmatrix}$$

利用初始向量 $x^{(0)}=(0, 0, 0)^T$，对下列每个方法执行二次迭代.

a. 雅可比　　b. 高斯-赛德尔　　c. 共轭梯度

计算机习题 4.7

1. 对共轭梯度法编写程序并进行测试. 一个好的测试实例是希尔伯特矩阵及一个简单的 b 向量：

$$a_{ij} = (i+j+1)^{-1} \qquad b_i = \frac{1}{3}\sum_{j=1}^{n} a_{ij} \qquad (1 \leqslant i,j \leqslant n)$$

2. 利用下列方法求解下面的方程组，初始为 $x^{(0)}=0$.

$$\begin{bmatrix} 10 & 1 & 2 & 3 & 4 \\ 1 & 9 & -1 & 2 & -3 \\ 2 & -1 & 7 & 3 & -5 \\ 3 & 2 & 3 & 12 & -1 \\ 4 & -3 & -5 & -1 & 15 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 12 \\ -27 \\ 14 \\ -17 \\ 12 \end{bmatrix}$$

a. 雅可比　　b. 高斯-赛德尔　　c. 共轭梯度

注：这里的系数阵是对称正定的但不是对角占优的.

4.8 高斯算法中的舍入误差分析

在本节中，我们将研究求解线性方程组

$$Ax = b \qquad (A \in \mathbb{R}^{n\times n}, x \in \mathbb{R}^n, b \in \mathbb{R}^n)$$

中不可避免地产生的舍入误差.

245

原来由 Wilkinson 给出的分析是对非行尺度主元的高斯算法的. 结果是误差的后验估计界形式. 从而，在计算的结论中可做出关于误差大小的断言.

在分析中，我们假定一开始就执行了适当的行交换，使得恰当的主元素总是位于所要求的位置. 行主元方法保证我们在算法的第 k 步 $|a_{kk}^{(k)}| \geqslant |a_{ik}^{(k)}|$，$i \geqslant k$（记号参照 4.3 节.），由刚才给出的不等式推出消元过程中所需要的乘子大小不会超过 1.

详细说明 LU 分解的公式如下：

$$a_{ij}^{(k+1)} = \begin{cases} a_{ij}^{(k)} & i \leqslant k \\ a_{ij}^{(k)} - \ell_{ik} a_{kj}^{(k)} & i > k \text{ 和 } j > k \\ 0 & j \leqslant k < i \end{cases}$$

$$\ell_{ik}=\begin{cases}0 & i<k\\ 1 & i=k\\ a_{ik}^{(k)}/a_{kk}^{(k)} & i>k\end{cases}$$

该过程以 $A^{(1)}=A$ 开始，以 $A^{(n)}=U$ 结束. 矩阵 L 和 U 分别是单位下三角阵和上三角阵.

在计算机中实际上产生的数用符号～加以区别，例如，$\tilde{\ell}_{ik}$是计算所得的数并存放在指派给 ℓ_{ik} 的存储位置. 所以符号～的意思是与机器相关的. 我们将假定在计算过程中没有上溢或下溢.

在描述计算机的实际计算中，我们使用 2.1 节中定义的函数 fl. 记得若 x 和 y 是机器数，$\odot$表示算术运算，则计算机将产生一个用 $\mathrm{fl}(x\odot y)$表示的机器数来代替 $x\odot y$. 由(8)式和习题 2.1.8，$x\odot y$ 和 $\mathrm{fl}(x\odot y)$之间的关系可用下式表示

$$\mathrm{fl}(x\odot y)=(x\odot y)(1-\delta)=(x\odot y)/(1-\delta') \tag{1}$$

其中 δ 和 δ' 是数量不超过单位舍入 ϵ 的数，ϵ 与所考虑的特定计算机有关.

把这些内容加以考虑时，高斯算法中计算的数可由下式描述

$$\tilde{a}_{ij}^{(k+1)}=\begin{cases}\tilde{a}_{ij}^{(k)} & i\leqslant k\\ \mathrm{fl}[\tilde{a}_{ij}^{(k)}-\mathrm{fl}(\tilde{\ell}_{ik}\tilde{a}_{kj}^{(k)})] & i>k,j>k\end{cases}$$

246

$$\tilde{\ell}_{ik}=\begin{cases}1 & i=k\\ \mathrm{fl}(\tilde{a}_{ik}^{(k)}/\tilde{a}_{kk}^{(k)}) & i>k\end{cases}$$

定理 1（$\tilde{L}\tilde{U}$ 分解定理） 设 A 是 $n\times n$ 非奇异矩阵，其元素是一台具有单位舍入 ϵ 的计算机中的机器数. 行主元高斯算法产生矩阵 $\tilde{L}$ 和 $\tilde{U}$ 使得

$$\tilde{L}\tilde{U}=A+E,\text{这里 } |e_{ij}|\leqslant 2n\epsilon \max_{1\leqslant i,j,k\leqslant n}|a_{ij}^{(k)}|$$

证明 引入量 δ_{ij} 和 δ'_{ij} 扮演(1)式中的 $\pm\delta$ 和 $\pm\delta'$ 的角色. 算法中计算的数服从下列等式

$$\tilde{a}_{ij}^{(k+1)}=[\tilde{a}_{ij}^{(k)}-(1-\delta_{ij})\tilde{\ell}_{ik}\tilde{a}_{kj}^{(k)}]/(1-\delta_{ij}') \qquad (i>k,j>k)$$

$$\tilde{\ell}_{ik}=(1+\delta_{ik})\tilde{a}_{ik}^{(k)}/\tilde{a}_{kk}^{(k)} \qquad (i>k)$$

这两个等式中的第 1 个也可以写成形式

$$\tilde{a}_{ij}^{(k+1)}=\tilde{a}_{ij}^{(k)}-\tilde{\ell}_{ik}\tilde{a}_{kj}^{(k)}+\delta_{ij}\tilde{\ell}_{ik}\tilde{a}_{kj}^{(k)}+\delta_{ij}'\tilde{a}_{ij}^{(k+1)} \qquad (i>k,j>k)$$

现在引进 $n\times n$ 矩阵 $\tilde{L}^{(k)}$，除了第 k 列对角元下面外其余元素均为 0：

$$\tilde{L}^{(k)}=\begin{bmatrix}0 & & & \\ & \ddots & & \\ & & 0 & \\ & & \tilde{\ell}_{k+1,k} & \\ & & \vdots & \ddots \\ & & \tilde{\ell}_{n,k} & & 0\end{bmatrix}$$

因而 $\tilde{A}^{(k)}-\tilde{L}^{(k)}\tilde{A}^{(k)}$ 是对 $\tilde{A}^{(k)}$ 应用某一个初等行运算的结果；即对 $k+1\leqslant i\leqslant n$，从第 i 行中减去$\tilde{\ell}_{ik}$乘第 k 行. 这差不多就是 $\tilde{A}^{(k+1)}$ 的定义. 其差别起因于舍入误差及在 $\tilde{A}^{(k+1)}$ 中我们取 $\tilde{a}_{ik}^{(k+1)}=0$，

$k+1\leqslant i\leqslant n$ 的事实. 所以可写成

$$\widetilde{A}^{(k+1)}=\widetilde{A}^{(k)}-\widetilde{L}^{(k)}\widetilde{A}^{(k)}+E^{(k)} \tag{2}$$

这里 $E^{(k)}$ 是误差补偿矩阵.

为了分析矩阵 $E^{(k)}$，首先考虑 $i>k$，$j=k$ 的情况. 从前面的等式，我们有

$$\begin{aligned}e_{ik}^{(k)}&=\tilde{a}_{ik}^{(k+1)}-a_{ik}^{(k)}+\tilde{\ell}_{ik}^{(k)}\tilde{a}_{kk}^{(k)}\\&=0-\tilde{a}_{ik}^{(k)}+\tilde{\ell}_{ik}\tilde{a}_{kk}^{(k)}\\&=-\tilde{a}_{ik}^{(k)}+(\tilde{a}_{ik}^{(k)}/\tilde{a}_{kk}^{(k)})(1+\delta_{ik})\tilde{a}_{kk}^{(k)}\\&=\delta_{ik}\tilde{a}_{ik}^{(k)}\quad(i>k)\end{aligned}$$

247

其次考虑 $i>k$，$j>k$ 的情况. 我们得到

$$\begin{aligned}e_{ij}^{(k)}&=\tilde{a}_{ij}^{(k+1)}-\tilde{a}_{ij}^{(k)}+\tilde{\ell}_{ik}^{(k)}\tilde{a}_{kj}^{(k)}\\&=\delta_{ij}\tilde{\ell}_{ik}\tilde{a}_{kj}^{(k)}+\delta'_{ij}\tilde{a}_{ij}^{(k+1)}\quad(i>k,j>k)\end{aligned}$$

$E^{(k)}$ 的其他元素全为 0. 现在把由(2)式所得的等式相加，$k=1,2,\cdots,n-1$. 结果可写成

$$\widetilde{L}^{(1)}\widetilde{A}^{(1)}+\cdots+\widetilde{L}^{(n-1)}\widetilde{A}^{(n-1)}+I\widetilde{A}^{(n)}=A^{(1)}+E^{(1)}+\cdots+E^{(n-1)}$$

注意 $\widetilde{L}^{(k)}\widetilde{A}^{(k)}$ 是一个矩阵，其行只不过是单个行向量

$$[\tilde{a}_{k1}^{(k)},\tilde{a}_{k2}^{(k)},\cdots,\tilde{a}_{kn}^{(k)}]$$

的各种各样倍数. 因为这个行向量与

$$[\tilde{a}_{k1}^{(n)},\tilde{a}_{k2}^{(n)},\cdots,\tilde{a}_{kn}^{(n)}]$$

是相同的，所以我们得出 $\widetilde{L}^{(k)}\widetilde{A}^{(k)}=\widetilde{L}^{(k)}\widetilde{A}^{(n)}$. 还记得 $A^{(1)}=A$ 以及所计算的上三角因子 U 是 $\widetilde{U}=\widetilde{A}^{(n)}$. 取 $E=\sum_{k=1}^{n-1}E^{(k)}$，我们有

$$(\widetilde{L}^{(1)}+\widetilde{L}^{(2)}+\cdots+\widetilde{L}^{(n-1)}+I)\widetilde{A}^{(n)}=A^{(1)}+E$$

或

$$\widetilde{L}\widetilde{U}=A+E$$

现在剩下的全部工作就是产生一个 $\|E\|$ 的界. 设 $\rho=\max\limits_{1\leqslant i,j,k\leqslant n}|A_{ij}^{(k)}|$. 因为预先的行交换保证了所有的乘子满足不等式 $|\tilde{\ell}_{ik}|\leqslant 1$. 于是，从前面给出的 $e_{ij}^{(k)}$ 的式子，我们有

$$|e_{ik}^{(k)}|=|\delta_{ik}||\tilde{a}_{ik}^{(k)}|\leqslant\varepsilon\rho\qquad(i>k)$$

$$|e_{ij}^{(k)}|=|\delta_{ij}\tilde{\ell}_{ik}\tilde{a}_{kj}^{(k)}+\delta'_{ij}\tilde{a}_{ij}^{(k+1)}|\leqslant 2\varepsilon\rho\quad(i>k,j>k)$$

由 E 的定义，立即可得

$$|e_{ij}|=\left|\sum_{k=1}^{n-1}e_{ij}^{(k)}\right|\leqslant\sum_{k=1}^{n-1}\left|e_{ij}^{(k)}\right|\leqslant 2n\varepsilon\rho$$ ■

为了在实际中应用这个定理，我们必须首先知道数

$$\rho=\max_{1\leqslant i,j\leqslant n}|a_{ij}^{(k)}|$$

这个数可以通过增加适当的代码，在算法的进程中计算.

248

定理 2(点积舍入误差定理) 若 $x_1,x_2,\cdots,x_n$ 和 $y_1,y_2,\cdots,y_n$ 是机器数，则以自然的方式算出

$$\sum_{i=1}^{n} x_i y_i$$

的机器数可表成 $\sum_{i=1}^{n} x_i y_i (1+\delta_i)$，其中 δ_i 满足 $|\delta_i| \leqslant 6(n+1)\varepsilon/5$.（数 ε 是机器的单位舍入误差并且我们假定 $n\varepsilon < 1/3$.）

证明 这个假设意味着计算将如下进行：

$$\begin{cases} z_0 = 0 \\ z_k = \mathrm{fl}[z_{k-1} + \mathrm{fl}(x_k y_k)] \quad (1 \leqslant k \leqslant n) \end{cases}$$

对 n 用归纳法证明 $|\delta_i| \leqslant (1+\varepsilon)^{n+2-i} - 1$. 对 $n=1$，我们有

$$z_1 = \mathrm{fl}(x_1 y_1) = x_1 y_1 (1+\delta_1) \qquad |\delta_1| \leqslant \varepsilon$$

此时，要证明的命题是 $|\delta_1| \leqslant (1+\varepsilon)^2 - 1$，因为 $\varepsilon \leqslant (1+\varepsilon)^2 - 1$，所以这个命题成立. 若定理对 $n=k-1$ 成立，则对 $n=k$，它的证明如下进行

$$\begin{aligned} z_k &= [z_{k-1} + x_k y_k (1+\delta')](1+\delta) \\ &= \Big[\sum_{i=1}^{k-1} x_i y_i (1+\delta_i) + x_k y_k (1+\delta')\Big](1+\delta) \\ &= \sum_{i=1}^{k-1} x_i y_i (1+\delta_i+\delta+\delta_i\delta) + x_k y_k (1+\delta+\delta'+\delta\delta') \end{aligned}$$

在前面的式子中，我们有下列这些界：

$$|\delta| \leqslant \varepsilon \qquad |\delta'| \leqslant \varepsilon \qquad |\delta_i| \leqslant (1+\varepsilon)^{k+1-i} - 1 \qquad (1 \leqslant i \leqslant k-1)$$

我们需要证明

$$|\delta_i + \delta + \delta_i\delta| \leqslant (1+\varepsilon)^{k+2-i} - 1 \qquad |\delta + \delta' + \delta\delta'| \leqslant (1+\varepsilon)^2 - 1$$

这些不等式中的第 1 个证明如下：

$$\begin{aligned} |\delta_i + \delta + \delta_i\delta| &\leqslant |\delta_i| + |\delta|(1+|\delta_i|) \\ &\leqslant (1+\varepsilon)^{k+1-i} - 1 + \varepsilon(1+\varepsilon)^{k+1-i} \\ &= (1+\varepsilon)^{k+2-i} - 1 \end{aligned}$$

[249] 以类似的方式证明第 2 个不等式. 归纳法证毕. 现在需要对 $k \leqslant n+1$ 作下列估算

$$\begin{aligned} (1+\varepsilon)^k - 1 &= [1 + k\varepsilon + \frac{1}{2}k(k-1)\varepsilon^2 + \cdots + \varepsilon^k] - 1 \\ &= k\varepsilon[1 + \frac{1}{2}(k-1)\varepsilon + \frac{1}{6}(k-1)(k-2)\varepsilon^2 + \cdots] \\ &\leqslant k\varepsilon[1 + \frac{1}{2}k\varepsilon + (\frac{1}{2}k\varepsilon)^2 + \cdots] \\ &= k\varepsilon/(1 - \frac{1}{2}k\varepsilon) < \frac{6}{5}k\varepsilon \end{aligned}$$

在最后一步中使用了假设 $k\varepsilon \leqslant n\varepsilon < 1/3$. 最后，我们有

$$|\delta_i| \leqslant (1+\varepsilon)^{n+2-i} - 1 \leqslant \frac{6}{5}(n+2-i)\varepsilon \leqslant \frac{6}{5}(n+1)\varepsilon$$

■

定理 3（扰动的单位下三角方程组定理） 设 L 是 $n \times n$ 单位下三角阵，其元素是机器数. 设 b 是 n 维向量，其分量是机器数. $Ly=b$ 的计算解是向量 $\tilde{y}$，它是下列方程的精确解：

$$(L+\Delta)\tilde{y}=b,\ |\Delta_{ij}|\leqslant\frac{6}{5}(n+1)\varepsilon|\ell_{ij}| \tag{3}$$

这里 ε 是机器的单位舍入误差，并假定 $n\varepsilon<1/3$.

证明 $Ly=b$ 的精确解是用下列公式计算的

$$y_i=b_i-\sum_{j=1}^{i-1}\ell_{ij}y_j \qquad (1\leqslant i\leqslant n)$$

计算值 $\tilde{y}_1$，$\tilde{y}_2$，…，$\tilde{y}_n$ 满足下列形式的等式

$$\tilde{y}_i=\left[b_i-\sum_{j=1}^{i-1}\ell_{ij}\tilde{y}_j(1+\delta_{ij})\right]\Big/(1+\delta_{ii}) \tag{4}$$

其中数 δ_{ij} 满足

$$|\delta_{ij}|\leqslant\frac{6}{5}(n+1)\varepsilon \qquad (1\leqslant j\leqslant i\leqslant n) \tag{5}$$

这个命题是 2.1 节中所作的假定和定理 1 的推论．(4)式可重新整理成

$$(1+\delta_{ii})\tilde{y}_i=b_i-\sum_{j=1}^{i-1}\ell_{ij}\tilde{y}_j(1+\delta_{ij})$$

再重新改写为

$$\sum_{j=1}^{i}\ell_{ij}\tilde{y}_j(1+\delta_{ij})=b_i$$

我们把这个式子理解为一个矩阵等式

$$(L+\Delta)\tilde{y}=b \tag{6}$$

250

其中 Δ 是下三角阵，其元素为 $\ell_{ij}\delta_{ij}$，$1\leqslant j\leqslant i\leqslant n$. 因此，

$$|\Delta_{ij}|=|\ell_{ij}||\delta_{ij}|\leqslant\frac{6}{5}(n+1)\varepsilon|\ell_{ij}|$$

■

定理 4(扰动上三角方程组定理) 设 U 是 $n\times n$ 上三角非奇异阵．若 U 和 c 的元素都是机器数且 $n\varepsilon<1/3$，则 $Uy=c$ 的计算解 $\tilde{y}$ 精确地满足下列扰动方程组：

$$(U+\Delta)\tilde{y}=c,\ |\Delta_{ij}|\leqslant\frac{6}{5}(n+1)\varepsilon|u_{ij}|$$

证明 这个证明留给读者作为习题 4.8.1. ■

定理 5(扰动方程组定理) 设 A 和 b 的元素都是机器数．若用行主元高斯算法解 $Ax=b$，则计算解 $\tilde{x}$ 是扰动方程组

$$(A+F)\tilde{x}=b,\text{其中 } |f_{ij}|\leqslant 10n^2\varepsilon\rho$$

的精确解．这里 n 是 A 的阶数，$\rho=\max\limits_{1\leqslant i,j,k\leqslant n}|a_{ij}^{(k)}|$，$\varepsilon$ 是机器的单位舍入误差，并假设 $n\varepsilon<1/3$.

证明 因为 $n\varepsilon<1/3$，由定理 1，3 和 4，我们有

$$A+E=\tilde{L}\tilde{U} \qquad |e_{ij}|\leqslant 2n\varepsilon\rho$$

$$(\tilde{L}+\Delta)\tilde{y}=b \qquad |\Delta_{ij}|\leqslant\frac{6}{5}(n+1)\varepsilon|\tilde{\ell}_{ij}|$$

$$(\tilde{U}+\Delta')\tilde{x}=\tilde{y} \qquad |\Delta'_{ij}|\leqslant\frac{6}{5}(n+1)\varepsilon|\tilde{u}_{ij}|$$

把这些等式结合在一起得到

$$\begin{aligned}b&=(\tilde{L}+\Delta)\tilde{y}=(\tilde{L}+\Delta)(\tilde{U}+\Delta')\tilde{x}=(\tilde{L}\tilde{U}+\Delta\tilde{U}+\tilde{L}\Delta'+\Delta\Delta')\tilde{x}\\&=(A+E+\Delta\tilde{U}+\tilde{L}\Delta'+\Delta\Delta')\tilde{x}=(A+F)\tilde{x}\end{aligned}$$

其中 $F=E+\Delta\tilde{U}+\tilde{L}\Delta'+\Delta\Delta'$. 为得到 F 的估计，我们利用上面给的界，并且注意到 $|\ell_{ij}|\leqslant 1$ 和

$$|u_{ij}|=|a_{ij}^{(n)}|\leqslant\rho$$

因此，

$$\begin{aligned}|f_{ij}|&\leqslant|e_{ij}|+\sum_{v=1}^{n}\{|\Delta_{iv}||\tilde{u}_{vj}|+|\tilde{\ell}_{iv}||\Delta'_{vj}|+|\Delta_{iv}||\Delta'_{vj}|\}\\&\leqslant 2n\varepsilon\rho+\frac{6}{5}n(n+1)\varepsilon\rho+\frac{6}{5}n(n+1)\varepsilon\rho+\frac{36}{25}n(n+1)^2\varepsilon^2\rho\\&=n^2\varepsilon\rho\left\{\frac{2}{n}+\frac{12}{5}\frac{n+1}{n}+\frac{36}{25}\varepsilon\left(\frac{n+1}{n}\right)^2\right\}\end{aligned}$$

251 花括号中的值不会超过 10. ■

通过仔细观察前面 5 个定理中的细节并用计算 $\|F\|_\infty$ 来代替，可以改进定理 5 中所给出的界. 读者可以查阅 Forsythe and Moler[1967]、Golub and van Loan[1989]、Wilkinson[1965]、Isaacson and Keller[1966]等有关著作.

增长因子

高斯消元法中 $n\times n$ 矩阵 A 的**增长因子**定义为

$$g_n(A)=\frac{\max_{i,j,k}|a_{ij}^{(k)}|}{\max_{i,j}|a_{i,j}|}$$

这里 $a_{ij}^{(k)}$ 表示在消元过程第 k 步后的 (i,j) 元素. 它作为用选主元高斯消元法数值求解 $Ax=b$ 中稳定性的度量，增长因子以相对方式度量数在消元过程中如何变大. 除非 $g_n(A)$ 很大，否则选主元的高斯消元法被认为是数值稳定的.

增长因子出现各种各样的界. Wilkinson[1965]建立了界

$$\frac{\|\tilde{x}-x\|_\infty}{\|x\|_\infty}\leqslant 4n^2g_n(A)\kappa_\infty(A)\varepsilon$$

这里 x 是 $Ax=b$ 的精确解，$\tilde{x}$ 是用部分选主元的高斯消元法[⊖]在一台具有相对机器精度 ε 的计算机上用浮点运算得到的计算解，$\kappa_\infty(A)$ 是用 ∞ 范数的 A 的条件数. 下面是另一个与向后误差有关的界，它类似于定理 5 中的界：

$$\frac{\|\tilde{A}-A\|_\infty}{\|A\|_\infty}<8n^3g_n(A)\varepsilon$$

这里计算解 $\tilde{x}$ 满足 $\tilde{A}\tilde{x}=b$. (例如，见 Golub and van Loan[1989].)

在部分选主元的高斯消元法中增长因子的一个明确上界是

$$g_n(A)\leqslant 2^{n-1}$$

在 Golub and Van Loan[1989]、Higham and Higham[1989]、Foster[1994]、Wilkinson[1965]的书中能找到达到这个上界的例子. Trefethen and Schreiber[1990]指出对随机矩阵 $g_n(A)$ 不会指数增长. 最近，Foster[1994]和 Wright[1993]给出了增长因子指数增长的例子. 在应用中，误差增长通常是十分小的. 因此，部分选主元的高斯消元法被认为是一个稳定的过程，并

252 且被完全有把握地广泛应用.

⊖ 这个选主元策略不同于我们的选法，因为它不涉及比例常数.

30 多年以前，Wilkinson 报告了在他的经验中即使用部分选主元，增长因子相当大地增加也是极其罕见的. Wilkinson 观察到寻求使 $g_n(A)>n$ 的矩阵是很困难的，并且他未发现自然地产生的因子大于 16 的例子. 鉴于 Wilkinson 的观察，他的注记被一些研究人员作为智力的挑战，并且努力寻求具有大的增长因子的实例. 例如，Dongarra，Bunch，Moler，and Stewart[1979]报告了一个例子，对部分选主元高斯消元法的增长因子 $g_n(A)$是 23.

对全主元高斯消元法，Wilkinson[1961]指出

$$g_n(A)\leqslant n^{1/2}[2\cdot 3^{1/2}\cdot 4^{1/3}\cdots n^{1/(n-1)}]^{1/2}$$

虽然，这个界是十分巨大的，但是 Wilkinson[1961]提及对大的 n，它比 2^n 小得多，并且证明显示真正的上界还要小得多. 而且，这个界是 n 的缓慢增长函数. 因为涉及额外的计算工作量，所以看来全主元与部分主元相比没有多大的实用价值.

Cryer[1968]发表了一个称为 **Wilkinson 猜测**的命题：若对 A 执行全主元高斯消元法，则 $g_n(A)\leqslant n$. Gould[1991]和 Edelman[1992] 对全主元找到一个 $g_n(A)> n$ 的例子，所以这个猜测被澄清是不成立的.

因为上面提到的这些例子很不简单，所以在许多情况下，找出这种例子是需要超级计算机或高级数学软件包的. 另一方面，某些专家思索高斯消元法中曾经忽略的大的误差增长因子的可能性. 最近，发现了导致部分选主元高斯消元法中灾难性的误差增长例子. 它确实是一个典型的值得进一步研究的问题. (例如，见 Edelman[1992]和 Foster[1994]的论文.)

习题 4.8

1. 证明定理 4.
2. 证明定理 1 中的不等式

$$\| E^{(k)}\|_\infty\leqslant[1+2(n-k)]\varepsilon\rho$$
$$\| E\|_\infty\leqslant n^2\varepsilon\rho$$

3. 证明定理 3 中的不等式

$$\|\Delta\|_\infty\leqslant\frac{3}{5}n(n+1)\ \varepsilon\max_{1\leqslant i,j\leqslant n}|\ell_{ij}|$$

253

第5章　数值线性代数精选

5.0　基本概念回顾

研究矩阵特征值需要很熟悉复数，我们将简要地回顾一下那些重要的基本概念. 熟知的读者可以跳过这些内容.

基本概念

实数域$\mathbb{R}$ 的不足之处是实系数的 n 次多项式不一定有 n 个实零点(或根). 例如多项式 $p(x)=x^2-2x+2$ 没有实零点. 克服这个不足之处的一种做法是扩充这个域使得它包含非实数元素 i. 元素 i 用等式 $i^2=-1$ 来刻画. 从而得到的这个新域用$\mathbb{C}$ 表示，而它的元素称为**复数**. 它们具有形式

$$\gamma=\alpha+\mathrm{i}\beta(\alpha,\beta \text{ 是实数})$$

γ 的**共轭**和**模**分别定义为

$$\bar{\gamma}=\alpha-\mathrm{i}\beta$$

$$|\gamma|=\sqrt{\alpha^2+\beta^2}$$

注意，$\bar{\bar{\gamma}}=\gamma$ 且 $\gamma\bar{\gamma}=|\gamma|^2$.

域$\mathbb{C}$ 就不具有曾经提及的$\mathbb{R}$ 的不足之处. 因此，我们有**代数基本定理**，它指出每个具有复系数的非常数多项式在复平面中至少有一个零点. 由此可得到每个 n 次多项式都可以表示成 n 个线性因子之积.

254

向量空间$\mathbb{C}^n$ 由全体复 n 维向量组成，例如

$$x=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$$

其中 $x_j\in\mathbb{C}$，$1\leqslant j\leqslant n$. 若复向量 x 用复数 λ 相乘，则结果是另一个复向量：

$$\lambda x=(\lambda x_1,\lambda x_2,\cdots,\lambda x_n)^{\mathrm{T}}$$

因此，我们可把$\mathbb{C}^n$ 看作纯量域$\mathbb{C}$ 上的一个向量空间，在空间$\mathbb{C}^n$ 中，**内积和欧几里得范数**分别定义为

$$\langle x,y\rangle=\sum_{j=1}^{n}x_j\bar{y}_j \quad 及 \quad \|x\|_2=\sqrt{\langle x,x\rangle}$$

注意到

$$\langle x,y\rangle=\overline{\langle y,x\rangle} \qquad \langle x,\lambda y\rangle=\bar{\lambda}\langle x,y\rangle$$

并且

$$\langle x+y,z\rangle=\langle x,z\rangle+\langle y,z\rangle$$

如果 A 是一个具有复元素的矩阵，那么 A^* 将表示其**共轭转置**：$(A^*)_{jk}=\bar{A}_{kj}$. 特别地，如果 x 是一个 $n\times1$ 矩阵(或列向量)，那么 $x^*=(\bar{x}_j)$是一个 $1\times n$ 矩阵(或行向量)，并且

$$y^*x=\langle x,y\rangle=\sum_{j=1}^{n}x_j\bar{y}_j$$

$$x^* x = \langle x, x \rangle = \| x \|_2^2 = \sum_{j=1}^{n} x_j \bar{x}_j = \sum_{j=1}^{n} | x_j |^2$$

设 A 是一个 $n \times n$ 矩阵(它的元素可能是复数). 设 λ 是一个纯量(一个复数). 如果方程

$$Ax = \lambda x \tag{1}$$

有一个非平凡解(即 $x \neq 0$), 那么 λ 是 A 的一个**特征值**. 而满足方程(1)的非零向量 x 是 A 对应于特征值 λ 的**特征向量**. 例如, 等式

$$\begin{bmatrix} 2 & 0 & 1 \\ 5 & -1 & 2 \\ -3 & 2 & -\frac{5}{4} \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ -4 \end{bmatrix} = -2 \begin{bmatrix} 1 \\ 3 \\ -4 \end{bmatrix}$$

告诉我们 -2 是已知的 3×3 矩阵的特征值而且 $(1, 3, -4)^T$ 是一个相应的特征向量. 注意, 255 一个特征向量的非零倍就是对应于同一个特征值的另一个特征向量.

方程(1)有一个非平凡解的条件等价于下列这些条件之一:

$$A - \lambda I \text{ 映射某个非零向量为 } 0 \text{ 向量} \tag{2}$$

$$A - \lambda I \text{ 奇异} \tag{3}$$

$$\det(A - \lambda I) = 0 \tag{4}$$

因此, 原则上我们可以解含 λ 未知值的方程(4), 从而计算 A 的特征值. 方程(4)称为矩阵 A 的**特征方程**. 如果把这个方程用更详细的方式写出, 那么有

$$\det \begin{bmatrix} a_{11} - \lambda & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} - \lambda & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} - \lambda \end{bmatrix} = 0$$

根据行列式定义, 函数为矩阵元素乘积的项之和, 方程(4)的左边为变量 λ 的 n 次多项式形式. 这个多项式是 A 的**特征多项式**. 倘若特征值作为特征方程的根, 它们的重数被计算在内, 我们立刻得到结论: 一个 $n\times n$ 矩阵恰好有 n 个特征值.

例 1 为了说明这些概念, 让我们求下列矩阵的特征值:

$$A = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 3 \\ 2 & 1 & 1 \end{bmatrix}$$

解 特征方程是

$$\det \begin{bmatrix} 1-\lambda & 2 & 1 \\ 0 & 1-\lambda & 3 \\ 2 & 1 & 1-\lambda \end{bmatrix} = -\lambda^3 + 3\lambda^2 + 2\lambda + 8$$

$$= -(\lambda - 4)(\lambda^2 + \lambda + 2) = 0$$

特征方程的根是

$$\lambda_1 = 4 \qquad \lambda_2 = -\frac{1}{2} + \frac{1}{2}\sqrt{7}\,\mathrm{i} \qquad \lambda_3 = -\frac{1}{2} - \frac{1}{2}\sqrt{7}\,\mathrm{i}$$

这说明实矩阵的特征值不一定是实数. ■

前面的过程是计算特征值的**直接法**. 对于小矩阵用手算，它可能是最好的方法. 而对大矩阵和自动计算，它不是着重推荐的. 排斥它的一个理由是多项式的根很可能是多项式系数的非常敏感的函数. 因此，系数中的任何误差(如舍入误差)就有可能导致数值求根中严重的不准确性. 对此，Wilkinson 给出的一个经典的例子已在 2.3 节的最后作了讨论. 256

5.1 矩阵特征值问题：幂法

5.1.1 幂法

我们下面将讨论的数值方法是幂法. 设计这个方法是为了计算主特征值及对应于这个主特征值的特征向量. 为顺利达到目的，理论上必须假定 A 有下列两个性质：

1. 存在单个模最大的特征值.

2. 存在 n 个线性无关的特征向量组.

按照第一个假定，特征值 λ_1，λ_2，…，λ_n 可列为

$$|\lambda_1| > |\lambda_2| \geqslant |\lambda_3| \geqslant \cdots \geqslant |\lambda_n|$$

按照第二个假定，存在$\mathbb{C}^n$ 的一个基$\{u^{(1)}, u^{(2)}, \cdots, u^{(n)}\}$使得

$$Au^{(j)} = \lambda_j u^{(j)} \qquad (1 \leqslant j \leqslant n) \tag{1}$$

设 $x^{(0)}$ 是$\mathbb{C}^n$ 的任意元素，使得当 $x^{(0)}$ 表示成基元素 $u^{(1)}$，$u^{(2)}$，…，$u^{(n)}$ 的线性组合时，$u^{(1)}$ 的系数不为 0. 因此

$$x^{(0)} = a_1u^{(1)} + a_2u^{(2)} + \cdots + a_nu^{(n)} \qquad (a_1 \neq 0) \tag{2}$$

然后，我们形成

$$x^{(1)} = Ax^{(0)} \qquad x^{(2)} = Ax^{(1)} \qquad \cdots \qquad x^{(k)} = Ax^{(k-1)}$$

因此

$$x^{(k)} = A^kx^{(0)} \tag{3}$$

不失一般性，在下面的分析中，与向量 $u^{(j)}$ 相乘的所有系数 a_j 不出现. 因此，我们可把等式(2)写成

$$x^{(0)} = u^{(1)} + u^{(2)} + \cdots + u^{(n)}$$

由这个等式和(3)式，我们有

$$x^{(k)} = A^ku^{(1)} + A^ku^{(2)} + \cdots + A^ku^{(n)}$$ 257

利用(1)式，我们得出

$$x^{(k)} = \lambda_1^ku^{(1)} + \lambda_2^ku^{(2)} + \cdots + \lambda_n^ku^{(n)}$$

最后一个等式可写成形式

$$x^{(k)} = \lambda_1^k\left[u^{(1)} + \left(\frac{\lambda_2}{\lambda_1}\right)^k u^{(2)} + \cdots + \left(\frac{\lambda_n}{\lambda_1}\right)^k u^{(n)}\right]$$

因为 $|\lambda_1| > |\lambda_j|$，$2 \leqslant j \leqslant n$，我们看到当 $k \to \infty$时，系数$(\lambda_j/\lambda_1)^k$ 趋向 0 并且括号中的向量收敛于 $u^{(1)}$.

为简化记号，我们把 $x^{(k)}$ 写成

$$x^{(k)} = \lambda_1^k[u^{(1)} + \varepsilon^{(k)}]$$

这里当 $k\to\infty$ 时，$\varepsilon^{(k)}\to 0$. 为了能够取比值，设 φ 是 $\mathbb{C}^n$ 上使得 $\varphi(u^{(1)})\neq 0$ 的任意的线性泛函. 记得线性泛函 φ 对纯量 α，β 和向量 x，y 满足 $\varphi(\alpha x+\beta y)=\alpha\varphi(x)+\beta\varphi(y)$.（例如，$\varphi$ 可简单地计算任何已知向量的第 j 个分量.）于是

$$\varphi(x^{(k)}) = \lambda_1^k[\varphi(u^{(1)}) + \varphi(\varepsilon^{(k)})] \tag{4}$$

因此，当 $k\to\infty$ 时，下列比值收敛于 λ_1：

$$r_k \equiv \frac{\varphi(x^{(k+1)})}{\varphi(x^{(k)})} = \lambda_1\left[\frac{\varphi(u^{(1)}) + \varphi(\varepsilon^{(k+1)})}{\varphi(u^{(1)}) + \varphi(\varepsilon^{(k)})}\right] \to \lambda_1$$

这就构成计算 λ_1 的幂法. 因为当 $k\to\infty$ 时，向量 $x^{(k)}$ 的方向越来越同 $u^{(1)}$ 成一条线，所以这个方法也能够给出特征向量 $u^{(1)}$. 在文献中可找到幂法的许多变形和改进.

5.1.2 算法

下面是刚才所述幂法的算法：

```
input n, A, x, M
output 0, x
for k=1 to M do
    y←Ax
    r←φ(y)/φ(x)
    x←y
    output k, x, r
end do
```

这里 φ 是某个线性泛函.

在幂法的实际实施中，引进规范化的向量 $x^{(k)}$ 是适当的，因为否则它们可能收敛到 0 或者 258 变成无界. 于是，我们可以修改迭代如下：

```
input n, A, x, M
output 0, x
for k=1 to M do
    y←Ax
    r←φ(y)/φ(x)
    x←y/‖y‖
    output k, x, r
end do
```

这里使用的范数可以是任何一种方便的范数，例如 ℓ_∞ 范数 $\|x\|_\infty=\max\limits_{1\leqslant j\leqslant n}|x_j|$. 而比值 r 与非规范形式算法中的那些比值相同.（习题 5.1.2 要求证明这个命题.）

例 1 对矩阵 $A=\begin{bmatrix}6 & 5 & -5\\ 2 & 6 & -2\\ 2 & 5 & -1\end{bmatrix}$ 和初始向量 $x=(-1, 1, 1)^{\mathrm{T}}$ 使用幂法.

解 用规范形式的算法编程，取线性泛函 φ 为 $\varphi(x)=x_2$. 下面对少数 k 的值给出规范化

向量 $x^{(k)}$ 和比值 r_k：

$$
\begin{array}{lll}
k=0 & x^{(0)}=(-1.000\,00,\ \ 1.000\,00,\ \ 1.000\,00) & \\
k=1 & x^{(1)}=(-1.000\,00,\ \ 0.333\,33,\ \ 0.333\,33) & r_0=\ \ 2.0 \\
k=2 & x^{(2)}=(-1.000\,00,-0.111\,11,-0.111\,11) & r_1=-2.0 \\
k=3 & x^{(3)}=(-1.000\,00,-0.407\,41,-0.407\,41) & r_2=\ \ 22.0 \\
k=4 & x^{(4)}=(-1.000\,00,-0.604\,94,-0.604\,94) & r_3=8.909\,1 \\
\vdots & \vdots & \vdots \\
k=6 & x^{(6)}=(-1.000\,00,-0.824\,42,\ \ 0.824\,42) & r_5=6.715\,08 \\
\vdots & \vdots & \vdots \\
k=28 & x^{(28)}=(-1.000\,00,-0.999\,98,-0.999\,98) & r_{27}=6.000\,07
\end{array}
$$

A 的主特征值是 6，而一个特征向量是$(1,\ 1,\ 1)^{\mathrm{T}}$. ■

当某些或全部特征值是复数时我们的幂法分析是成立的，并且在我们的假设下算法计算 λ_1 和一个相应的特征向量.

5.1.3 艾特肯加速

如果把比值 r_k 看作是 λ_1 的近似，那么估计 $|r_k-\lambda_1|$ 的误差是有趣的. 这里有关的结果都包含在习题 5.1.3 和 5.1.4 中. 结果是

$$r_{k+1}-\lambda_1=(c+\delta_k)(r_k-\lambda_1)$$

259

常数 c 满足 $|c|<1$，数 δ_k 收敛于 0. 按 1.3 节中的术语，这就推得序列$[r_k]$线性收敛于 λ_1. 因为我们知道误差的这个线性收敛行为，所以可使用一个称为**艾特肯加速**的一般过程. 从已知的序列$[r_k]$出发，我们利用公式

$$s_k=\frac{r_kr_{k+2}-r_{k+1}^2}{r_{k+2}-2r_{k+1}+r_k}$$

构造另一个序列$[s_k]$. 依照下面一般的结果，这个序列收敛快于原来的序列.

定理 1（艾特肯加速定理） *设$[r_n]$是一个收敛于极限 r 的序列. 如果 $r_{n+1}-r=(c+\delta_n)(r_n-r)$，$|c|<1$ 且$\lim\limits_{n\to\infty}\delta_n=0$，则新序列*

$$s_n=\frac{r_nr_{n+2}-r_{n+1}^2}{r_{n+2}-2r_{n+1}+r_n}\quad(n\geqslant 0)$$

收敛于 r 快于$[r_n]$. 实际上，当 $n\to\infty$时，$(s_n-r)/(r_n-r)\to 0$.

证明 定义误差序列 $h_n\equiv r_n-r$. 则一个简短的计算可得

$$
\begin{aligned}
s_n&=\frac{(r+h_n)(r+h_{n+2})-(r+h_{n+1})^2}{(r+h_{n+2})-2(r+h_{n+1})+(r+h_n)}\\
&=r+\frac{h_nh_{n+2}-h_{n+1}^2}{h_{n+2}-2h_{n+1}+h_n}
\end{aligned}
$$

利用假设 $h_{n+1}=(c+\delta_n)h_n$得到 $h_{n+2}=(c+\delta_{n+1})(c+\delta_n)h_n$以及

$$
\begin{aligned}
s_n-r&=\frac{h_n(c+\delta_{n+1})(c+\delta_n)h_n-(c+\delta_n)^2h_n^2}{(c+\delta_{n+1})(c+\delta_n)h_n-2(c+\delta_n)h_n+h_n}\\
&=h_n\frac{(c+\delta_{n+1})(c+\delta_n)-(c+\delta_n)^2}{(c+\delta_{n+1})(c+\delta_n)-2(c+\delta_n)+1}
\end{aligned}
\tag{5}
$$

因为等式(5)中分子收敛于 0 而分母收敛于$(c-1)^2$且不等于 0，所以显然有$\lim\limits_{n\to\infty}(s_n-r)/h_n=0$. ■

因为公式中减法相消将最终损害结果，所以重要的是在明显地出现稳定值后，立刻终止艾特肯加速过程.

在讨论幂法的其他变形之前，我们将证明一个关于特征值的基本事实.

定理 2(逆矩阵的特征值定理) 若λ是A的一个特征值且A是非奇异的，则λ^{-1}是A^{-1}的一个特征值.

260

证明 设$Ax=\lambda x$，$x\neq 0$，则$x=A^{-1}(\lambda x)=\lambda A^{-1}x$. 因此，$A^{-1}x=\lambda^{-1}x$并且$\lambda^{-1}$是$A^{-1}$的一个特征值. ■

5.1.4 逆幂法

定理 2 提出一种计算A的最小特征值的方法. 假如A的特征值可以排列如下：

$$|\lambda_1|\geqslant|\lambda_2|\geqslant\cdots\geqslant|\lambda_{n-1}|>|\lambda_n|>0$$

这个假设推出A是非奇异的，因为 0 不是一个特征值. A^{-1}的特征值是λ_j^{-1}，并且它们被安排成这样：

$$|\lambda_n^{-1}|>|\lambda_{n-1}^{-1}|\geqslant\cdots\geqslant|\lambda_1^{-1}|>0$$

因此，我们可以对A^{-1}应用幂法来计算λ_n^{-1}. 首先计算A^{-1}，再计算$x^{(k+1)}=A^{-1}x^{(k)}$不是一个好的主意. 我们宁可通过解方程

$$Ax^{(k+1)}=x^{(k)}$$

得到$x^{(k+1)}$. 这可以用高斯消元法有效实施. 它只要执行一次高斯消元法的分解过程，然后从$x^{(0)}$起到$x^{(1)}$再到$x^{(2)}$等等改变右端向量反复执行求解过程. 这就是**逆幂法**.

例 2 用例 1 中的矩阵说明逆幂法. 它的LU分解是

$$\begin{bmatrix}6&5&-5\\2&6&-2\\2&5&-1\end{bmatrix}=\begin{bmatrix}1&0&0\\\frac{1}{3}&1&0\\\frac{1}{3}&\frac{10}{13}&1\end{bmatrix}\begin{bmatrix}6&5&-5\\0&\frac{13}{3}&-\frac{1}{3}\\0&0&\frac{12}{13}\end{bmatrix}$$

解 我们从向量$x=(3,7,-13)^{\mathrm{T}}$开始并计算 25 步. 在每一步，我们解$Ux^{(k+1)}=L^{-1}x^{(k)}$得到$x^{(k+1)}$. 再计算和打印比值，即$r_k=x_1^{(k+1)}/x_1^{(k)}$. 在我们进行下一步之前，规范化$x^{(k+1)}$(即用它的$\ell_\infty$范数除它). 下面是一些输出数据：

$k=0$ $x^{(0)}=(\ 3.000\,00,\ 7.000\,00,-13.000\,00)$

$k=1$ $x^{(1)}=(-0.801\,65,-0.008\,26,-1.000\,00)$ $r_0=-5.888\,9$

$k=2$ $x^{(2)}=(-0.950\,89,-0.017\,74,-1.000\,00)$ $r_1=1.197\,59$

$k=3$ $x^{(3)}=(-0.987\,59,-0.007\,12,-1.000\,00)$ $r_2=1.027\,50$

$k=4$ $x^{(4)}=(-0.996\,88,-0.002\,23,-1.000\,00)$ $r_3=1.004\,46$

⋮ ⋮ ⋮

$k=6$ $x^{(6)}=(-0.999\,80,-0.000\,17\ -1.000\,00)$ $r_5=1.000\,12$

⋮ ⋮ ⋮

$k=11$ $x^{(11)}=(-1.000\,00,\ 0.000\,00,-1.000\,00)$ $r_{10}=1.000\,00$

其余的迭代没有改变. ■ [261]

在这个阶段，我们已经概述了计算 A 的最大特征值的**幂法**和计算 A 的最小特征值的**逆幂法**. 通过考虑**位移矩阵** $A-\mu I$，我们能得到计算 A 最靠近一个给定值 μ 的特征值的方法. 假定 A 的一个特征值 λ_k 满足不等式 $0<|\lambda_k-\mu|<\varepsilon$，这里 μ 是一个预先指定的复数. 假如 A 的其余特征值都满足不等式 $|\lambda_j-\mu|>\varepsilon$. 因为 $A-\mu I$ 的特征值是 $\lambda_j-\mu$，所以可以对 $A-\mu I$ 应用逆幂法，结果为 $(\lambda_k-\mu)^{-1}$ 的一个近似值. 这里再从解方程 $(A-\mu I)x^{(k+1)}=x^{(k)}$ 得到所需要的向量序列. 在这个算法中计算一次 $A-\mu I$ 的高斯分解. 因为这个过程计算 $z=(\lambda_k-\mu)^{-1}$，所以我们可以从公式 $\lambda_k=z^{-1}+\mu$ 得到 λ_k. 这个算法称为**位移逆幂法**.

类似地，我们可以计算一个离给定值 μ 最远的特征值 λ_k. 假如对 A 的某个特征值 λ_k，存在一个正的 ε 使得 $|\lambda_k-\mu|>\varepsilon$，并且对 A 的其他一切特征值 λ_j，我们有 $0<|\lambda_j-\mu|<\varepsilon$. 对 $(A-\mu I)$ 应用幂法计算 $z=\lambda_k-\mu$. 因此，我们得到 $\lambda_k=z+\mu$.

5.1.5 小结

我们把这些结果总结在下表中：

方法	方程	计算
幂	$x^{(k+1)}=Ax^{(k)}$	最大特征值 λ_1
逆幂	$Ax^{(k+1)}=x^{(k)}$	最小特征值 λ_n
位移幂	$x^{(k+1)}=(A-\mu I)x^{(k)}$	离 μ 最远的特征值
位移逆幂	$(A-\mu I)x^{(k+1)}=x^{(k)}$	离 μ 最近的特征值

习题 5.1

1. 设 A 是一个 $n\times n$ 矩阵，并且其有 n 个线性无关的特征向量 $\{u^{(1)}, u^{(2)}, \cdots, u^{(n)}\}$. 设 $Au^{(i)}=\lambda_i u^{(i)}$，并设 P 是由向量 $u^{(1)}, u^{(2)}, \cdots, u^{(n)}$ 为其列的矩阵. 试问 $P^{-1}AP$ 是什么？
2. 证明：若幂法的规范化和非规范化形式以同样的初始向量开始，则两个算法中 r 的值相同.
3. 在幂法中，设 $r_k=\varphi(x^{(k+1)})/\varphi(x^{(k)})$. 我们知道 $\lim\limits_{k\to\infty}r_k=\lambda_1$. 证明相对误差服从
$$\frac{r_k-\lambda_1}{\lambda_1}=\left(\frac{\lambda_2}{\lambda_1}\right)^k c_k$$
这里数 c_k 形成一个有界序列.
4. (续) 证明 $r_{k+1}-\lambda_1=(c+\delta_k)(r_k-\lambda_1)$，$|c|<1$ 且 $\lim\limits_{k\to\infty}\delta_k=0$，故可以应用艾特肯加速. 假设 $|\lambda_2|>|\lambda_3|$.
5. 倘若 $c\neq 0$，证明在艾特肯加速中，当 $n\to\infty$ 时，$(s_n-r)/(r_{n+2}-r)\to 0$.
6. 计算基本(非规范化)幂法执行 m 步所需要的乘法和/或除法次数. [262]
7. 在规范化幂法中，证明：若 $\lambda_1>0$，则向量 $x^{(k)}$ 收敛于一个特征向量.
8. 设计一个简单的幂法乘法处理下列情况：$\lambda_1=-\lambda_2>|\lambda_3|\geqslant|\lambda_4|\geqslant\cdots\geqslant|\lambda_n|$.
9. 如果序列 $[r_n]$ 只满足假设：$|r_{n+1}-r|\leqslant c|r_n-r|$，$0<c<0.2$，你对艾特肯加速能证明什么？
10. 设 A 的特征值满足 $\lambda_1>\lambda_2>\cdots>\lambda_n$(全是实数，但不一定是正的). 为了使幂法应用于 $A+\beta I$ 时能最迅速地收敛于 λ_1，参数 β 应该使用什么值？
11. 如果不是 A 就是 B 为非奇异阵，证明 $I-AB$ 和 $I-BA$ 有相同的特征值.
12. 如果对一个实矩阵与一个实初始向量应用幂法. 当主特征值是复数时将会发生什么情况？怎样应用课本中

所概述的理论？

13. 下列矩阵的特征多项式是怎样的？

$$\begin{bmatrix} a_1 & a_2 & a_3 & \cdots & a_{n-1} & a_n \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

14. 对任意的复数 λ 和任意的 $n\times n$ 矩阵 A，证明

$$\dim\{x: Ax = \lambda x\} = n - \operatorname{rank}(A-\lambda I)$$

15. 设 $A=LU$，其中 L 是单位下三角阵而 U 是上三角阵. 取 $B=UL$，证明 B 和 A 有相同的特征值.

16. 假如 A 有一行，比如说第 k 行，使得 $\sum_{j=1,j\neq k}^{n} |a_{kj}|=0$. 设 B 表示从 A 中去掉第 k 行和第 k 列后所得到的矩阵. 证明 a_{kk} 是 A 的特征值，并且 A 的其余特征值都是 B 的特征值.

17. 一个 $n\times n$ 矩阵称为亏损的，如果它的特征向量不张成 $\mathbb{R}^n$. 证明矩阵 $A=\begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}$ 是亏损的.

18. 证明：如果一个 $n\times n$ 矩阵有 n 个不同的特征值，则它不是亏损的.

19. (续)证明上题的逆命题不成立.

20. 在一些计算实矩阵特征值的实验中，时常观察到出现实特征值. 如果 n 是奇数，证明一个 $n\times n$ 实矩阵至少一定有一个实特征值.

21. (续)一个实多项式可分解成二次因式和一次因式. 二次因式可能有复根. 计算实二次因式 x^2+ax+b 有虚根的概率，假定系数 a 和 b 选自正方形 $|a|\leqslant\rho$，$|b|\leqslant\rho$ 中均匀分布的随机变量. 当 $\rho\to\infty$ 时，证明这个概率收敛到 0，当 $\rho\to 0$ 时，证明这个概率收敛于 1/2. 这个结论对实矩阵的特征值暗示着什么？

22. 对大的 k 值计算 $x^{(k)}=A^k x$，我们可以反复执行矩阵的平方：

263

$$A\to A^2\to A^4\to A^8\to\cdots\to A^{2m}$$

因而一个单独的矩阵向量积产生 $x^{(2^m)}=A^{2^m}x$. 把这个过程与通常的幂法进行比较. 计算出每种方法达到 $x^{(2^m)}$ 所需要的乘法次数. 证明：对大的 m，平方法总是比较经济的. 再通过调节 A 的逐次乘幂，设计一个避免上溢的方法.

23. 使用 φ 的无穷范数，采取幂法的两种迭代确定下列矩阵的谱半径 $\rho(A)$ 的近似值. 初始向量取 $(1, 1, 1)^T$. (范数不是线性泛函.)

$$A=\begin{bmatrix} 2 & 0 & -1\\ -2 & -10 & 0\\ -1 & -1 & 4\end{bmatrix}$$

24. 证明：当 φ 是范数时，幂法中的比值 r_k 收敛于 $|\lambda_1|$.

25. 证明计算艾特肯加速的一个合适形式是

$$s_k = r_k - \frac{(\Delta r_k)^2}{\Delta^2 r_k}$$

其中

$$\Delta r_k = r_{k+1}-r_k \qquad \Delta^2 r_k = \Delta r_{k+1}-\Delta r_k$$

这些是第 6 章中将要讨论的向前差分.

26. 对范围 $1\leqslant j, k\leqslant n$ 中的每个 j 和 k，设 f_{jk} 是 λ 的线性函数(一个线性函数具有形式 $\lambda\to\alpha\lambda+\beta$.) 你如何确定 λ 的值使 $\det(f_{jk}(\lambda))=0$？就一般而言，存在多少这样的值？它们位于复平面的什么地方？

计算机习题 5.1

1. 对例 1 中的矩阵作幂法，初始向量为$(1, 2, 3)^T$. 做 100 步，并且解释为什么在早期貌似收敛之后又收敛到另一个值.
2. 对已知的 $n\times n$ 矩阵 A，编写一个子程序或过程，应用正规的幂法 M 步，从一个已知的向量 x 开始. 结合艾特肯加速. 在每步，都打印当前的向量 $x^{(k)}$，当前的比值 r_k，以及课本中所定义的当前加速值 s_k. 对 $A-\mu I$测试这个过程，其中

a. $A=\begin{bmatrix}5 & 4 & 1 & 1\\ 4 & 5 & 1 & 1\\ 1 & 1 & 4 & 2\\ 1 & 1 & 2 & 4\end{bmatrix}$，$\mu=0$，3，6，11

b. $A=\begin{bmatrix}2 & 3 & 4\\ 7 & -1 & 3\\ 1 & -1 & 5\end{bmatrix}$，$\mu=0$，5，10

3. 编写一个逆幂法的计算机程序，并且选择一些矩阵来测试这个程序.
4. a. 编写一个计算机程序，在你的计算机系统上重新产生例 1 中所给的结果. 把你的程序模块化为若干个子程序或过程，计算算法的每个主要部分(例如，你可以要求构造下列子程序：(1)一个矩阵乘一个向量，(2)点积，(3)用另一个向量替代一个向量，(4)向量的范数，(5)规范化一个向量以及其他项目.

   264

   b. 把艾特肯加速增加到你的代码中并再次运行它. 对这个方案写出你的结论.
5. a. 对例 2 重复上题 a.

   b. 对例 2 重复上题 b.
6. 编写并测试一个计算最远离给定复数的特征值的计算机程序. 并对本节中的矩阵例子测试程序.
7. 构造一个例子说明如果不在一个适当的步数中止，艾特肯加速将会产生毫无意义的结果.

5.2 舒尔定理和 Gershgorin 定理

我们继续复习基本的矩阵特征值理论. 记得如果存在一个非奇异阵 P 使得 $B=PAP^{-1}$，就称两个矩阵 A 和 B 是互为**相似的**. 这个概念的重要性源自用不同的基底表示同一个线性变换的两个矩阵是互为相似的这个定理.

定理 1(相似矩阵的特征值定理) 相似矩阵有相同的特征值.

证明 设 A 和 B 是相似矩阵，即

$$B=PAP^{-1}$$

我们将看到 A 和 B 有相同的特征多项式，事实上

$$\begin{aligned}\det(B-\lambda I)&=\det(PAP^{-1}-\lambda I)\\&=\det[P(A-\lambda I)P^{-1}]\\&=\det P\det(A-\lambda I)\det P^{-1}\\&=\det(A-\lambda I)\end{aligned}$$

这里我们引用了两个基本事实：两个矩阵积的行列式是它们的行列式之积，逆矩阵的行列式是这个矩阵行列式的倒数. ■

5.2.1 舒尔分解

定理1提出了一个求 A 的特征值的方法：用一个相似变换把 A 变为 B，即 $B=PAP^{-1}$，并计算 B 的特征值. 如果 B 比 A 简单，那么其特征值的计算就可能比较容易. 特别是，假如 B 为三角阵. 则 B 的特征值(且 A 的特征值)就是 B 的对角元. 因为前面的方法(在理论上)总是可能的，据此自然地就引导我们得到重要的舒尔定理. 记得如果 $UU^*=I$，就称矩阵 U 为**酉阵**，这里 U^* 表示 U 的共轭转置：$(U^*)_{ij}=\overline{U}_{ji}$. 如果对某个酉阵 U 有 $B=UAU^*$，那么称矩阵 A 和 B 是**酉相似**的.

265

定理2(舒尔定理) 每个方阵酉相似于一个三角阵.

证明 对矩阵 A 的阶数 n 作归纳法. 对 $n=1$，定理是平凡的. 现在假设该定理对所有 $n-1$ 阶矩阵都成立，然后考虑 n 阶矩阵 A. 设 λ 是 A 的一个特征值，x 是一个相应的特征向量，不失一般性地，假设 $\|x\|_2=1$. 如往常一样，设 $e^{(1)}$ 表示标准的单位向量 $e^{(1)}=(1, 0, \cdots, 0)^T$. 如果 $x_1\neq 0$，设 $\beta=x_1/|x_1|$；如果 $x_1=0$，取 $\beta=1$. 由引理2可知，存在一个酉阵 U 使得 $Ux=\beta e^{(1)}$. 因为 U 是酉阵，所以 $U^{-1}=U^*$ 且 $\beta^{-1}x=U^*e^{(1)}$，故

$$UAU^*e^{(1)}=UA\beta^{-1}x=\beta^{-1}\lambda Ux=\lambda e^{(1)}$$

这就证明了 UAU^* 的第一列是 $\lambda e^{(1)}$. 设 $\widetilde{A}$ 表示从 UAU^* 中删除第一行和第一列后所得到的矩阵. 由归纳假设，存在一个 $n-1$ 阶酉阵 Q 使得 $Q\widetilde{A}Q^*$ 是三角阵. 因此化 A 为三角阵的酉阵是

$$V=\begin{bmatrix}1 & 0\\ 0 & Q\end{bmatrix}U$$

因为

$$\begin{aligned}VAV^* &= \begin{bmatrix}1 & 0\\ 0 & Q\end{bmatrix}UAU^*\begin{bmatrix}1 & 0\\ 0 & Q^*\end{bmatrix}\\ &= \begin{bmatrix}1 & 0\\ 0 & Q\end{bmatrix}\begin{bmatrix}\lambda & w\\ 0 & \widetilde{A}\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & Q^*\end{bmatrix}\\ &= \begin{bmatrix}1 & 0\\ 0 & Q\end{bmatrix}\begin{bmatrix}\lambda & wQ^*\\ 0 & \widetilde{A}Q^*\end{bmatrix}=\begin{bmatrix}\lambda & wQ^*\\ 0 & Q\widetilde{A}Q^*\end{bmatrix}\end{aligned}$$

在上面等式中，w 是 $n-1$ 维行向量，0表示 $n-1$ 维零向量. 由于矩阵

$$\begin{bmatrix}1 & 0\\ 0 & Q\end{bmatrix}$$

是酉阵，所以证毕. ■

读者可以看到刚才所给出的证明并不是构造性的，这是因为它要求特征值存在(并利用它们)，但是几乎没有提到如何计算它们的问题.

266

推论1(相似矩阵推论) 每个方阵相似于一个三角阵.

推论2(酉相似矩阵推论) 每个埃尔米特阵酉相似于一个对角阵.

证明 若 A 是埃尔米特阵，则 $A=A^*$. 设 U 是一个酉阵并使得 UAU^* 是上三角的，则

$(UAU^*)^*$是下三角的. 但是

$$(UAU^*)^* = U^{**}A^*U^* = UAU^*$$

因此，矩阵UAU^*既是上三角的又是下三角的，所以它一定是对角阵. ■

引理 1(酉阵第一引理) 矩阵$I-vv^*$是酉阵当且仅当$\|v\|_2^2=2$或$v=0$.

证明 为使矩阵$U=I-vv^*$是酉阵，我们要求

$$\begin{aligned} I &= UU^* \\ &= (I-vv^*)(I-vv^*) \\ &= I-2vv^*+vv^*vv^* \\ &= I-2vv^*+(v^*v)(vv^*) \\ &= I-(2-v^*v)vv^* \end{aligned}$$

(这里我们把纯量v^*v移到前面.) 因此，对v的充分必要条件就是$v^*v=2$或$vv^*=0$. ■

引理 2(酉阵第二引理) 设x和y是两个使得$\|x\|_2=\|y\|_2$且$\langle x, y\rangle$为实数的向量，则存在一个形如$I-vv^*$的酉阵，使得$Ux=y$.

证明 若$x=y$，设$v=0$. 若$x\neq y$，一个非常好的猜想引导我们去尝试$v=\alpha(x-y)$，$\alpha=\sqrt{2}/\|x-y\|_2$. 如此选择的$v$导致

$$\begin{aligned} Ux-y &= (I-vv^*)x-y = x-vv^*x-y \\ &= x-y-\alpha^2(x-y)(x^*-y^*)x \\ &= (x-y)[1-\alpha^2(x^*x-y^*x)] \end{aligned}$$

因为上式方括号中的数为0，所以其结果变成0. 为此，我们利用假设$x^*x=y^*y$及$y^*x=x^*y$计算

$$\begin{aligned} 1-\alpha^2(x^*x-y^*x) &= 1-\frac{1}{2}\alpha^2(x^*x+x^*x-y^*x-y^*x) \\ &= 1-\frac{1}{2}\alpha^2(x^*x+y^*y-y^*x-x^*y) \\ &= 1-\frac{1}{2}\alpha^2(x^*-y^*)(x-y) \\ &= 1-\frac{1}{2}\alpha^2\|x-y\|_2^2 = 0 \end{aligned}$$

■ 267

例 1 下面是舒尔分解$UAU^*=T$的一个具体例子，其中U是酉阵，T为上三角阵.

解

$$\begin{bmatrix} 0.36 & 0.48 & 0.80 \\ 0.48 & 0.64 & -0.60 \\ 0.80 & -0.60 & 0.00 \end{bmatrix}\begin{bmatrix} 361 & 123 & -180 \\ 148 & 414 & -240 \\ -92 & 169 & 65 \end{bmatrix}\begin{bmatrix} 0.36 & 0.48 & 0.80 \\ 0.48 & 0.64 & -0.60 \\ 0.80 & -0.60 & 0.00 \end{bmatrix}$$

$$= \begin{bmatrix} 125 & 380 & -125 \\ 0 & 465 & 1\,250 \\ 0 & 0 & 250 \end{bmatrix}$$ ■

如果已知$n\times n$矩阵A的一个特征值λ，那么舒尔定理的证明指出，怎样产生一个$(n-1)\times(n-1)$

矩阵$\widetilde{A}$，其特征值除λ外都与A的那些特征值相同．这个过程称为**降阶**.

特征值**降阶过程**的形式如下：

1. 相应一个已知的特征值λ得到一个特征向量x.
2. 若$x_1 \neq 0$，定义$\beta = x_1/|x_1|$，否则取$\beta=1$.
3. 定义$\alpha=\sqrt{2}/\|x-\beta e^{(1)}\|_2$，$v=\alpha(x-\beta e^{(1)})$，$U=I-vv^*$.
4. 从UAU^*中去掉第一行和第一列后得到矩阵$\widetilde{A}$.

求多项式p的根的一个类似的降阶过程在3.5节中作过讨论．在求出一个根ξ之后，我们可以用$x-\xi$除$p(x)$得到一个具有相同根(除ξ外)的低次多项式.

大多数计算特征值的数值方法一次只计算一个特征值．而把任意一种这样的方法与降阶过程组合起来就可以计算出所希望多的矩阵特征值．但实际上，必须细心地使用这个方法，因为后续特征值可能会受到越来越多的舍入误差影响.

5.2.2 特征值的定位

文献中的许多定理都以一种粗略的方式描述了位于复平面上的矩阵特征值．下面是这些定位定理中最著名的定理：

定理3(Gershgorin定理)　下列复平面中n个圆盘D_i的并

$$D_i = \{z \in \mathbb{C} : |z - a_{ii}| \leqslant \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|\} \qquad (1 \leqslant i \leqslant n)$$

268 包含$n\times n$矩阵A的谱(即它的特征值集).

证明　设λ是A的谱中任意元．选择一个向量x使得$Ax=\lambda x$且$\|x\|_\infty=1$．设i是使$|x_i|=1$的一个指标．因为$(Ax)_i=\lambda x_i$，我们有

$$\lambda x_i = \sum_{j=1}^{n} a_{ij} x_j$$

所以，

$$(\lambda - a_{ii}) x_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij} x_j$$

再取绝对值并用三角不等式和$|x_j| \leqslant 1 = |x_i|$，有

$$|\lambda - a_{ii}| \leqslant \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}||x_j| \leqslant \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|$$

于是，$\lambda \in D_i$. ■

例2　矩阵

$$A = \begin{bmatrix} -1+\mathrm{i} & 0 & \frac{1}{4} \\ \frac{1}{4} & 1 & \frac{1}{4} \\ 1 & 1 & 3 \end{bmatrix}$$

269 的Gershgorin圆盘如图5-1所示．从此图可以推知A的所有特征值满足不等式$1/2 \leqslant |\lambda| \leqslant 5$.

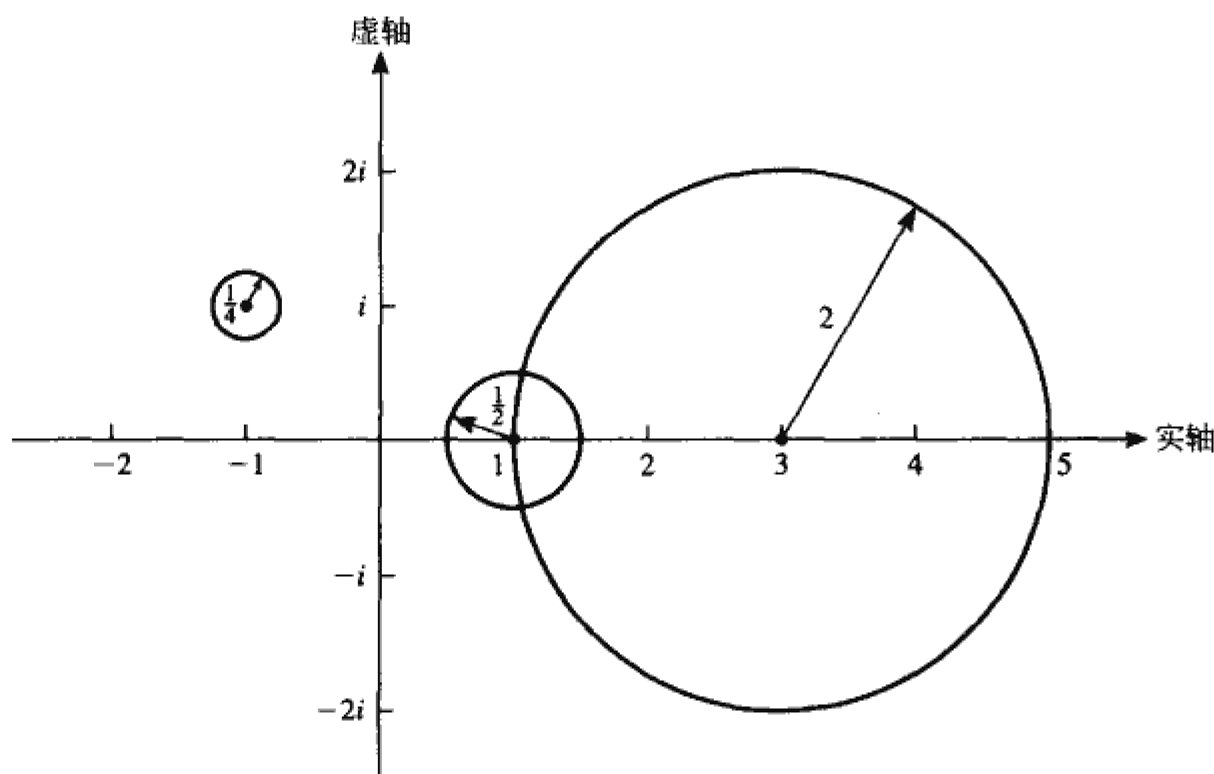


图 5-1 Gershgorin 圆盘 ■

定理 4(特征值圆盘定理) *若矩阵 A 用相似变换 $P^{-1}AP$ 对角化，而 B 是任意矩阵，则 $A+B$ 的特征值位于下列圆盘之并中：*

$$\{\lambda\in\mathbb{C}:|\lambda-\lambda_i|\leqslant\kappa_\infty(P)\ \|B\|_\infty\}$$

其中 λ_1，λ_2，…，λ_n 是 A 的特征值，$\kappa_\infty(P)$ 是 4.4 节中所定义的 P 的条件数.

证明 我们可以建立一个稍加精确的结果．若 $P^{-1}AP=D$．对角阵 D 的对角元为 λ_1，λ_2，…，λ_n．用"sp"表示谱，我们有

$$\mathrm{sp}(A+B)=\mathrm{sp}[P^{-1}(A+B)P]=\mathrm{sp}(D+P^{-1}BP)$$

对 $D+C$ 应用 Gershgorin 定理，$C=P^{-1}BP$，可以推出 $A+B$ 的谱在 $D+C$ 的 Gershgorin 圆盘的并中，即

$$\left\{\lambda\in\mathbb{C}:|\lambda-\lambda_i-c_{ii}|\leqslant\sum_{\substack{j=1\\j\neq i}}^{n}|d_{ij}+c_{ij}|=\sum_{\substack{j=1\\j\neq i}}^{n}|c_{ij}|\right\}$$

由三角不等式和 $\|C\|_\infty$ 的定义，我们有

$$\begin{aligned}|\lambda-\lambda_i|&\leqslant|\lambda-\lambda_i-c_{ii}|+|c_{ii}|\leqslant|c_{ii}|+\sum_{\substack{j=1\\j\neq i}}^{n}|c_{ij}|\\&\leqslant\|C\|_\infty\leqslant\|P^{-1}\|_\infty\|B\|_\infty\|P\|_\infty=\kappa_\infty(P)\|B\|_\infty\end{aligned}$$ ■

定理 4 指出当 A 被扰动后，其特征值将被一个不超过 $\kappa_\infty(P)\|B\|_\infty$ 的量扰动，其中 B 是扰动矩阵而 $\kappa_\infty(P)$ 是用 ℓ_∞ 矩阵范数计算的 P 的条件数．(见 4.4 节.)

对埃尔米特阵(即 $A^*=A$)，矩阵 P 可选为酉阵；这是舒尔定理的推论 2．因为，此时 P 的行就是满足 $\|x\|_2=1$ 的向量．由此可得 $\|P\|_\infty\leqslant\sqrt{n}$．这对 P^{-1} 也成立，所以 $\kappa_\infty(P)\leqslant n$．因此，对任何矩阵 B，$A+B$ 的特征值位于下列圆盘的并中：

$$\{\lambda \in \mathbb{C}: |\lambda-\lambda_i| \leqslant n\|B\|_\infty\}$$

其中 λ_1，λ_2，…，λ_n 是埃尔米特阵 A 的特征值.

习题 5.2

1. 求下列矩阵的舒尔分解：

270

$$\begin{bmatrix} 3 & 8 \\ -2 & 3 \end{bmatrix} \text{和} \begin{bmatrix} 4 & 7 \\ 1 & 12 \end{bmatrix}$$

2. 设 $\|x\|_2=1$，取 $U=I-2xx^*$. 证明 $U^2=I$.
3. (续)若 $x^*x=2$，那么 $(I-xx^*)^{-1}$ 是什么？
4. (续)设 $x^*x=1$，确定 $I-xx^*$ 是否可逆？
5. (续)证明 $I-xx^*$ 奇异当且仅当 $x^*x=1$. 并在所有非奇异情况下，求出它的逆.
6. 证明 A 的特征值位于下面所定义的两个集合 D 和 E 的交中.

$$D=\bigcup_{i=1}^{n}\{z\in\mathbb{C}: |z-a_{ii}| \leqslant \sum_{\substack{j=1\\j\neq i}}^{n}|a_{ij}|\}$$

$$E=\bigcup_{i=1}^{n}\{z\in\mathbb{C}: |z-a_{ii}| \leqslant \sum_{\substack{j=1\\j\neq i}}^{n}|a_{ji}|\}$$

7. 证明：若 λ 是 A 的一个特征值，则存在一个非零向量 x 使得 $x^TA=\lambda x^T$. (其中 x^T 表示行向量.)
8. 证明：若 A 是埃尔米特阵，则课本中的降阶方法将会产生一个埃尔米特阵.
9. 证明：若矩阵 A 的一列，比如说第 j 列，满足 $a_{ij}=0$，$i\neq j$，则 a_{jj} 是 A 的一个特征值.
10. **规范阵**是一个与其伴随矩阵可交换的矩阵：$AA^*=A^*A$. 证明：若 A 是规范阵，则对任意的纯量 λ，$A-\lambda I$ 也是规范阵.
11. 假如 A 是规范阵且 x 和 y 是 A 对应于不同特征值的特征向量. 证明 $x^*y=0$.
12. 证明：若 A 是规范阵，则 A 和 A^* 有相同的特征向量.
13. 证明：若 A 是规范阵，则由条件 $AB=BA$ 可推出 $A^*B=BA^*$.
14. 证明：若 x 和 y 是 $\mathbb{C}^n$ 中具有相同欧几里得范数的点，则存在一个酉阵 U 使得 $Ux=y$.
15. 证明或否定：若 $\{x_1, x_2, \cdots, x_k\}$ 和 $\{y_1, y_2, \cdots, y_k\}$ 是 $\mathbb{C}^n$ 中的标准正交集，则存在一个酉阵 U 使得 $Ux_i=y_i$，$1\leqslant i\leqslant k$.
16. 证明：若对三个向量 v，x，y，有 $(I-vv^*)x=y$，则 $\langle x, y\rangle$ 是实数.
17. 为使 $I-uv^*$ 是酉阵，求向量对 u 和 v 应达到的确切条件.
18. 证明：若 Q 是酉阵，则

$$\|A\|_2 = \|QA\|_2 = \|AQ\|_2$$

注：$\|A\|_2^2=\rho(AA^T)$.

19. 证明：若 A 是对角阵，则

$$\|A\|_2 = \max_{1\leqslant i\leqslant n}|a_{ii}|$$

20. 证明：对任何方阵 A，$\|A\|_2^2\leqslant\|A^*A\|_2$.
21. 设 A_j 表示 A 的第 j 列. 证明：$\|A_j\|_2\leqslant\|A\|_2$. 请问这个结论对所有从属矩阵范数是否成立？

271

22. 矩阵 A 的迹是 $\mathrm{tr}(A)=\sum_{i=1}^{n}a_{ii}$，证明矩阵 A 的迹等于其特征值之和.(此处使用舒尔定理会有帮助.)
23. (续)证明：若 λ_1，λ_2，…，λ_n 是 A 的特征值，则 A^m 的迹是

$$\mathrm{tr}(A^m)=\lambda_1^m+\lambda_2^m+\cdots+\lambda_n^m$$

24.(续)证明：若 A 的特征值满足 $|\lambda_1|>|\lambda_i|$，$i=2,3,\cdots,n$，则

$$\lambda_1=\lim_{m\to\infty}\frac{\mathrm{tr}(A^{m+1})}{\mathrm{tr}(A^m)}$$

25. 证明：若 A 和 B 是方阵，则 AB 和 BA 有相同的特征值.

26. 证明：若 $a^2+b^2=1$，$c^2+d^2=1$，则下列矩阵是酉阵：

$$\begin{bmatrix} ad & ac & b \\ bd & bc & -a \\ c & -d & 0 \end{bmatrix}$$

注意：对任意的 θ 和 φ，我们可设 $a=\sin\theta$，$b=\cos\theta$，$c=\sin\varphi$，$d=\cos\varphi$.

27. 证明：若向量 x 满足 $\|x\|_2=1$，则存在一个酉阵，其第一列是 x.

28. 设 A 是 $n\times n$ 矩阵，B 是 $m\times m$ 矩阵，C 是 $n\times m$ 矩阵. 证明：若 C 的秩为 m 并且 $AC=CB$，则

$$\mathrm{sp}(B)\subseteq\mathrm{sp}(A)$$

29. 证明或否定：若 A 是方阵，则存在一个埃尔米特酉阵 U 使得 UAU 是三角阵.

30. 不用求出特征值，证明矩阵

$$A=\begin{bmatrix} 6 & 2 & 1 \\ 1 & -5 & 0 \\ 2 & 1 & 4 \end{bmatrix}$$

的特征值满足不等式 $1\leqslant|\lambda|\leqslant 9$.

31. 证明

$$\begin{bmatrix} 3 & \frac{1}{3} & \frac{2}{3} \\ 1 & -4 & 0 \\ \frac{1}{2} & \frac{1}{2} & -1 \end{bmatrix}$$

的特征值的虚部都位于区间$[-1, 1]$中.

32. 求矩阵

$$A=\begin{bmatrix} 2.888 & 0.984 & -1.440 \\ 1.184 & 3.312 & -1.920 \\ -0.160 & 2.120 & -0.200 \end{bmatrix}$$

的舒尔分解 $UAU^*=T$. 提示：这里应该使用例 1 中的酉阵.

33. 若 $|a_{ii}-\lambda|>\sum_{j=1,j\neq i}^{n}|a_{ij}|$，则矩阵 $A-\lambda I$ 是对角占优的. 利用这个概念证明 Gershgorin 定理.

34. 设 $U=I-\lambda uu^*$，其中 u 是一个给定的向量. 试求使 U 为酉阵的一切复数 λ.

35. 假如矩阵 A 相似于一个几乎对角阵 B. 那么能否推出 A 的特征值接近于 B 的对角元？证明一个与此有关的定理，并且研究当 B 是几乎三角阵时的问题.

272

36. 设 A 是 $n\times n$ 矩阵并设 $D_1, D_2, \cdots, D_n$ 是其 Gershgorin 圆盘. 假如特征值 λ 位于 D_k 中但不在任何其他圆盘中. 设 x 是对应于 λ 的一个特征向量. 证明：对 $i\neq k$，$|x_k|>|x_i|$.

37. a. 对矩阵

$$A=\begin{bmatrix} 0 & 2 & -1 \\ -2 & -10 & 0 \\ -1 & -1 & 4 \end{bmatrix}$$

概略画出 Gershgorin 圆盘并给出谱半径 $\rho(A)$的一个界.

b. 利用 $\|A\|_1$ 确定 $\rho(A)$的上界和下界，其中

$$A=\begin{bmatrix}1 & 0 & 1\\ 1 & 4 & 0\\ 0 & 0 & 3\end{bmatrix}$$

之后再重复使用 Gershgorin 定理.

38. 利用 Gershgorin 定理证明因为对角占优阵的特征值不为 0，所以它是非奇异阵.

39. 证明：$\det(I+xx^*)=1+x^*x$. 提示：存在一个酉阵把 x 映射成 $e_1=(1, 0, \cdots, 0)^T$ 的倍向量.

5.3 正交分解和最小二乘问题

本节我们继续利用复向量和矩阵来进行研究. 在本章开始，我们回顾了复空间$\mathbb{C}^n$ 中的基本代数. 在此空间中，允许我们用内积$\langle , \rangle$定义正交的概念. $\mathbb{C}^n$ 中的一组向量$[v_1, v_2, \cdots, v_k]$被称为**正交的**，如果当 $i\neq j$ 时，就有$\langle v_i, v_j\rangle=0$. 若$\langle v_i, v_j\rangle=\delta_{ij}$，则这组向量被称为**标准正交的**. 取 $i=j$，我们看到对每个 i 可以推出 $\|v_i\|_2=1$. 若 v_i 是一个 $n\times k$ 矩阵 A 的列，则这个标准正交性可以用等式 $A^*A=I$ 来表示.

5.3.1 基本概念

现设$[v_1, v_2, \cdots, v_n]$是$\mathbb{C}^n$ 的标准正交基. 每个元 $x\in\mathbb{C}^n$ 对适当的复数 c_i 具有下列唯一的表达式

$$x=\sum_{i=1}^{n}c_iv_i$$

在这个等式的两边对某个 v_j 取内积，可得到

273

$$\langle x,v_j\rangle=\left\langle\sum_{i=1}^{n}c_iv_i,v_j\right\rangle=\sum_{i=1}^{n}c_i\langle v_i,v_j\rangle=\sum_{i=1}^{n}c_i\delta_{ij}=c_j$$

这确保对一切 $x\in\mathbb{C}^n$，

$$x=\sum_{i=1}^{n}\langle x,v_i\rangle v_i \tag{1}$$

成立. 项$\langle x, v_i\rangle v_i$ 表示 x 在 v_i 方向的**分量**. $\mathbb{R}^2$ 中典型情况的草图参见图 5-2.

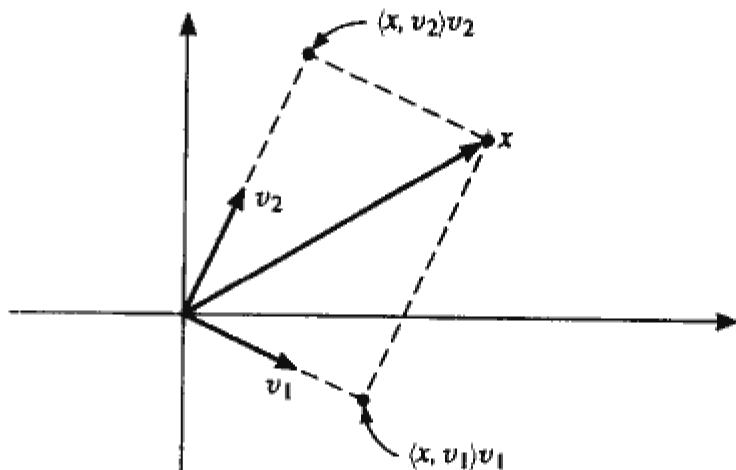


图 5-2 向量的正交分量

内积空间是一个抽象的代数系统，$\mathbb{C}^n$ 是其中一个具体代表. 它是一个(复数域上)线性空间，在空间内定义的内积满足下列公理：

1. 若 $x\neq 0$，则 $\langle x, x\rangle >0$.
2. $\langle \alpha x+\beta y, z\rangle=\alpha\langle x, z\rangle+\beta\langle y, z\rangle$，其中 α，$\beta\in\mathbb{C}$.
3. $\langle x, y\rangle=\overline{\langle y, x\rangle}$.

按$\mathbb{C}^n$ 中的方式给出范数、正交和标准正交的定义. 内积空间不一定就是有限的. 由前面的公理很容易证明$\langle x, y+z\rangle=\langle x, y\rangle+\langle x, z\rangle$和$\langle x, \alpha y\rangle=\bar{\alpha}\langle x, y\rangle$.

在任何内积空间中**毕达哥拉斯法则**都是成立的：若$\langle x, y\rangle=0$，则

$$\| x+y\|_2^2 = \| x\|_2^2+ \| y\|_2^2$$

这是从

$$\| x+y\|_2^2 = \langle x+y,x+y\rangle = \langle x,x\rangle + \langle y,x\rangle + \langle x,y\rangle + \langle y,y\rangle = \| x\|_2^2+ \| y\|_2^2$$

中得出的.

5.3.2 格拉姆-施密特过程

古典的**格拉姆-施密特过程**可用来得到任何内积空间中的标准正交系. 为描述它，我们假定在内积空间中给定一个线性无关的向量序列$[x_1, x_2, \cdots]$.（这个序列可以是有限的或无限的.）我们用公式 274

$$u_k = \| x_k - \sum_{i<k}\langle x_k,u_i\rangle u_i \|_2^{-1}\left(x_k - \sum_{i<k}\langle x_k,u_i\rangle u_i\right) \qquad (k\geqslant 1) \tag{2}$$

生成一个标准正交序列$[u_1, u_2, \cdots]$.

定理 1(格拉姆-施密特序列定理) *格拉姆-施密特序列$[u_1, u_2, \cdots]$具有性质：对 $k\geqslant 1$，$\{u_1, u_2, \cdots, u_k\}$是$\{x_1, x_2, \cdots, x_k\}$线性生成空间的标准正交基.*

证明 我们对 k 作数学归纳法. 对 $k=1$，我们从上面(2)式有 $u_1=\| x_1\|_2^{-1}x_1$. 因此，集$\{u_1\}$是标准正交的，并且它的线性生成空间与$\{x_1\}$的线性生成空间是一样的. 我们注意到集$\{x_1, x_2, \cdots\}$是线性无关的，所以$\| x_1\|_2\neq 0$.

由归纳假设，假如$\{u_1, u_2, \cdots, u_{k-1}\}$是 $\mathrm{span}\{x_1, x_2, \cdots, x_{k-1}\}$的标准正交基. 设

$$v = x_k - \sum_{i<k}\langle x_k,u_i\rangle u_i \tag{3}$$

由于，对 $j<k$，

$$\begin{aligned}\langle v,u_j\rangle &= \langle x_k,u_j\rangle - \sum_{i<k}\langle x_k,u_i\rangle\langle u_i,u_j\rangle \\ &= \langle x_k,u_j\rangle - \sum_{i<k}\langle x_k,u_i\rangle\delta_{ij} = \langle x_k,u_j\rangle - \langle x_k,u_j\rangle = 0\end{aligned}$$

因此 v 正交于 $u_1, u_2, \cdots, u_{k-1}$. 若 $v=0$，我们从(3)式看到 $x_k\in\mathrm{span}\{u_1, u_2, \cdots, u_{k-1}\}$. 由归纳假设，我们得到 $x_k\in\mathrm{span}\{x_1, x_2, \cdots, x_{k-1}\}$，这与假定$\{x_1, x_2, \cdots, x_k\}$线性无关相矛盾. 因此，$v\neq 0$，且 u_k 由表达式$(\| v\|_2)^{-1}v$ 完全确定. 因为 u_k 的范数等于 1，所以集合$\{u_1, u_2, \cdots, u_k\}$是标准正交的. 归纳假设连同(3)式一起表明 v(及 u_k)在 $\mathrm{span}\{x_1, x_2, \cdots, x_k\}$中，所以，$\mathrm{span}\{u_1, u_2, \cdots, u_k\}\subseteq\mathrm{span}\{x_1, x_2, \cdots, x_k\}$. 又因为$\{u_1, u_2, \cdots, u_k\}$和$\{x_1, x_2, \cdots, x_k\}$都是线性无关的(两者都生成 k 维空间)，所以 $\mathrm{span}\{u_1, u_2, \cdots, u_k\}=\mathrm{span}$

$\{x_1, x_2, \cdots, x_k\}$. ■

若对矩阵的列用格拉姆-施密特过程，那么我们可以说明这个结果为一个矩阵的分解．此时，在计算中产生的内积可以存放在矩阵中，它将是因子之一．我们对 $m\times n$ 矩阵 A 的列 A_1，A_2，…，A_n 应用这个过程，n 步之后即可得到一个 $m\times n$ 矩阵 B，其列构成一个标准正交集．下面就是**格拉姆-施密特算法**：

275

```
for j=1 to n do
    for i=1 to j-1 do
        t_ij ← ⟨A_j, B_i⟩
    end do
    C_j ← A_j − Σ_{i<j} t_ij B_i
    t_jj ← ‖C_j‖_2
    B_j ← t_jj^{-1} C_j
end do
```

定理 2（格拉姆-施密特分解定理）　当对秩为 n 的 $m\times n$ 矩阵 A 应用格拉姆-施密特过程时，会产生一个分解

$$A = BT \tag{4}$$

其中 B 是一个具有标准正交列的 $m\times n$ 矩阵，而 T 是一个具有正对角元的 $n\times n$ 上三角阵．

证明　首先，观察前面的算法实际上是执行(2)式中压缩的格拉姆-施密特过程．其次，当 $i>j$ 时，取 $t_{ij}=0$ 来完善 T 矩阵的定义．由定理 1，B 的列构成$\mathbb{R}^m$ 中 n 个向量的一个标准正交集，并且每个 A_j 都是 B_1，B_2，…，B_j 的一个线性组合．事实上，由(1)式，

$$\begin{aligned} A_j &= \sum_{i=1}^{j}\langle A_j, B_i\rangle B_i = \sum_{i=1}^{j-1}\langle A_j, B_i\rangle B_i + \langle A_j, B_j\rangle B_j \\ &= \sum_{i=1}^{j-1} t_{ij}B_i + \langle A_j, B_j\rangle B_j \end{aligned} \tag{5}$$

现在，我们计算

$$\begin{aligned} \langle A_j, B_j\rangle &= \left\langle C_j + \sum_{i<j} t_{ij}B_i, B_j \right\rangle = \langle C_j, B_j\rangle \\ &= \langle C_j, C_j\rangle / t_{jj} = t_{jj} \end{aligned} \tag{6}$$

当(6)式的结果被应用于(5)式时，得到

$$A_j = \sum_{i=1}^{j} t_{ij}B_i = \sum_{i=1}^{n} t_{ij}B_i \qquad (1 \leqslant j \leqslant n) \tag{7}$$

这是(4)式的另外一种表示方式． ■

5.3.3 修正的格拉姆-施密特算法

经验表明(Rice[1966])，格拉姆-施密特过程的某种重组通常有较好的数值性质．下面是**修正的格拉姆-施密特算法**：

276

```
for k=1 to n do
    A_k ← (‖A_k‖_2)^{-1} A_k
    for j=k+1 to n do
```

```
        A_j ← A_j − ⟨A_j, A_k⟩A_k
    end do
end do
```

其中 A_j 是矩阵 A 的第 j 列. 可以看出从 A_j 的分量中尽快地同时去掉了基向量 A_k 方向的分量. 在这个算法中，有相当多的覆盖，结束时，原来的集$\{A_1, A_2, \cdots, A_n\}$已被一个标准正交集替代.

为了避免计算 $\|x\|_2$ 时涉及到平方根，修正的格拉姆-施密特算法通常给出如下形式，它会产生一个稍微不同的分解：

```
for k=1 to n do
    d_k ← ‖A_k‖₂²
    t_kk ← 1
    for j=k+1 to n do
        t_kj ← d_k⁻¹⟨A_j, A_k⟩
        A_j ← A_j − t_kj A_k
    end do
end do
```

定理 3(修正的格拉姆-施密特分解定理) *若对秩为 n 的 $m\times n$ 矩阵 A 的列应用修正的格拉姆-施密特过程，则变换后的 $m\times n$ 矩阵 B 具有列的正交集且满足*

$$A = BT$$

其中 T 是一个 $n\times n$ 单位上三角阵，它的元素 $t_{kj}(j>k)$ 是在算法中生成的.

证明 为证明此定理，我们用不是覆盖量而是保存量的这种方式来编写算法. 如果不这样做，那么在证明中 A_j 就会不确定，因为在算法的不同阶段，它表示不同的向量. 我们把原来的集重新标记成$\{A_1^{(1)}, A_2^{(1)}, \cdots, A_n^{(1)}\}$. 现在这个算法可写为：

```
for k=1 to n do
    d_k ← ‖A_k^(k)‖₂²
    for j=k+1 to n do
        t_kj ← d_k⁻¹⟨A_j^(k), A_k^(k)⟩
        A_j^(k+1) ← A_j^(k) − t_kj A_k^(k)
    end do
    for j=1 to k do
        A_j^(k+1) ← A_j^(k)
    end do
end do
```

277

(在这个算法中不存在覆盖.)

下列的命题将对 k 用归纳法证明：

$$\mathcal{A}(k): 若 \min(i,j) < k, 则 \langle A_i^{(k)}, A_j^{(k)}\rangle = d_i\delta_{ij}.$$

注意$\mathcal{A}(1)$是成立的，因为没有满足 $\min(i, j)<1$ 的数对. 用$\mathcal{A}(k)$作为归纳假设. 则$\mathcal{A}(k+1)$的证明如下进行：考虑任何满足 $\min(i, j)<k+1$ 的数对(i, j). 由对称性，我们可假设$i\leqslant j$. 那么有 4 种情况需要分析.

1. 若 $i<k$ 和 $j\leqslant k$，则由$\mathcal{A}(k)$，得

$$\langle A_i^{(k+1)}, A_j^{(k+1)} \rangle = \langle A_i^{(k)}, A_j^{(k)} \rangle = d_i\delta_{ij}$$

2. 若 $i<k$ 和 $j>k$，则由 $\mathcal{A}(k)$，得

$$\begin{aligned}\langle A_i^{(k+1)}, A_j^{(k+1)} \rangle &= \langle A_i^{(k)}, A_j^{(k)} - t_{kj}A_k^{(k)} \rangle \\ &= \langle A_i^{(k)}, A_j^{(k)} \rangle - \bar{t}_{kj}\langle A_i^{(k)}, A_k^{(k)} \rangle \\ &= d_i\delta_{ij} - \bar{t}_{kj}d_i\delta_{ik} = 0\end{aligned}$$

3. 设 $i=j=k$，则有

$$\langle A_k^{(k+1)}, A_k^{(k+1)} \rangle = \langle A_k^{(k)}, A_k^{(k)} \rangle = d_k$$

4. 设 $i=k<j$，则有

$$\begin{aligned}\langle A_k^{(k+1)}, A_j^{(k+1)} \rangle &= \langle A_k^{(k)}, A_j^{(k)} - t_{kj}A_k^{(k)} \rangle = \langle A_k^{(k)}, A_j^{(k)} \rangle - \bar{t}_{kj}\langle A_k^{(k)}, A_k^{(k)} \rangle \\ &= \bar{t}_{kj}d_k - \bar{t}_{kj}d_k = 0\end{aligned}$$

由上述归纳证明，$\mathcal{A}(n)$成立，这意味着当 $i<n$ 时，$\langle A_i^{(n)}, A_j^{(n)} \rangle = d_i\delta_{ij}$. 因而集合$\{A_1^{(n)}, A_2^{(n)}, \cdots, A_n^{(n)}\}$是正交的. 若矩阵 B 的第 j 列是 $A_j^{(n)}$，则 $B^*B=\text{diag}(d_1, d_2, \cdots, d_n)$. 现在我们取 $t_{kk}=1$，并且当 $k>j$ 时，取 $t_{kj}=0$. 要使得 $A=BT$，注意 A 的第 k 列由下式给出.

$$\begin{aligned}A_k = A_k^{(1)} &= (A_k^{(1)} - A_k^{(2)}) + (A_k^{(2)} - A_k^{(3)}) + \cdots + (A_k^{(k-1)} - A_k^{(k)}) + A_k^{(k)} \\ &= t_{1k}A_1^{(1)} + t_{2k}A_2^{(2)} + \cdots + t_{k-1,k}A_{k-1}^{(k-1)} + t_{kk}A_k^{(k)} \\ &= t_{1k}A_1^{(n)} + t_{2k}A_2^{(n)} + \cdots + t_{kk}A_k^{(n)}\end{aligned}$$

278 ■

5.3.4 最小二乘问题

这里所讨论的正交因子分解的最重要应用就是线性方程组的最小二乘问题. 考察下列 n 个未知数 m 个方程的方程组

$$Ax = b \tag{8}$$

其中矩阵 A 是 $m\times n$，x 是 $n\times 1$，b 是 $m\times 1$. 我们假定 A 的秩为 n；因此 $m\geqslant n$. 因为 b 平常不位于由 A 的列生成的$\mathbb{C}^m$ 的子空间内，所以方程组(8)通常无解. 此时，经常需要求一个使得残差向量 $b-Ax$ 范数最小的 x. (8)式的最小二乘"解"是使 $\|b-Ax\|_2$ 极小的向量 x.（在关于 A 的秩的假设下，这个 x 是唯一的.）

引理 1（最小二乘问题引理） *若 x 是一个使 $A^*(Ax-b)=0$ 的点，则 x 是最小二乘问题的解.*

证明 设 y 是其他任何点. 因为 $A^*(Ax-b)=0$，所以 $b-Ax$ 正交于 A 的列空间. 而且，由于 $A(x-y)$在 A 的列空间内，所以我们有$\langle b-Ax, A(x-y)\rangle=0$，由毕达哥达斯法则可证

$$\begin{aligned}\|b-Ay\|_2^2 &= \|b-Ax+A(x-y)\|_2^2 \\ &= \|b-Ax\|_2^2 + \|A(x-y)\|_2^2 \geqslant \|b-Ax\|_2^2\end{aligned}$$

■

若 A 被分解成上述定理所描述的形式 $A=BT$，则方程组 $Ax=b$ 的最小二乘解将是下列 $n\times n$ 方程组的精确解

$$Tx = (B^*B)^{-1}B^*b$$

用引理 1 验证这个结果如下：

$$A^*Ax = (BT)^*BTx = T^*B^*B(B^*B)^{-1}B^*b = T^*B^*b = A^*b$$

矩阵$(B^*B)^{-1}$是 $\text{diag}(d_1^{-1}, d_2^{-1}, \cdots, d_n^{-1})$，数 d_i 是修正的格拉姆-施密特算法中所计算的那

些值．如前所述，这样的安排避免了平方根的计算．

与方程组 $Ax=b$ 有关的最小二乘问题的另一个方法是直接使用引理 1．于是，若

$$A^*(Ax-b)=0$$

则量 $\|b-Ax\|_2$ 达到极小．若 $m\times n$ 矩阵 A 的秩为 n，则 A^*A 是 $n\times n$ 非奇异矩阵(习题 5.3.2)，并且，此时恰好存在一个最小二乘解；通过解 $n\times n$ 非奇异**正规方程组**

$$A^*Ax=A^*b \tag{9}$$

279

它被唯一地确定．矩阵 A^*A 是埃尔米特正定的(习题 5.3.3)．所以可以用楚列斯基分解求解(9)式．当 A 的秩小于 n 时，方程(9)相容但它可能有许多解．

由于正规方程概念上的简单性，直接使用正规方程求解最小二乘问题是非常吸引人的．然而，它又被认为是解决这个问题令人最不满意的方法之一．如此判断的一个原因是因为 A^*A 的条件数可能要比 A 的条件数坏得多．用下例来说明这一现象：

$$A=\begin{bmatrix}1 & 1 & 1\\ \varepsilon & 0 & 0\\ 0 & \varepsilon & 0\\ 0 & 0 & \varepsilon\end{bmatrix}\qquad A^*A=\begin{bmatrix}1+\varepsilon^2 & 1 & 1\\ 1 & 1+\varepsilon^2 & 1\\ 1 & 1 & 1+\varepsilon^2\end{bmatrix} \tag{10}$$

在 A 中非零参数 ε 都是为了防止 A 的秩为 1．而在 A^*A 中，防止矩阵秩为 1 的参数仅仅是 ε^2．于是，A^*A 的非奇异性更要慎重处理小的 ε．当然，在一台计算机中，可能会有秩等于 3 的 A 和秩等于 1 的 A^*A．

5.3.5 豪斯霍尔德 QR 分解

我们现在转到最有用的正交因子分解，它是由 Alston Householder(豪斯霍尔德)给出并以他的名字命名的．目标是把 $m\times n$ 矩阵 A 分解成因子乘积

$$A=QR$$

其中 Q 为 $m\times m$ 酉阵，R 为 $m\times n$ 上三角阵．因子分解算法实际上产生了

$$Q^*A=R$$

而 Q^* 是由具有特别形式

$$\begin{bmatrix}I_k & 0\\ 0 & I_{m-k}-vv^*\end{bmatrix}$$

的酉阵乘积逐步组成的．这些矩阵称为**反射**或**豪斯霍尔德变换**．关于矩阵 A 的秩我们不作任何假设．

首先，我们确定 $v\in\mathbb{C}^n$ 使得 $I-vv^*$ 是酉阵并且使得 $(I-vv^*)A$ 开始看上去像 R．特别是，它的第 1 列应该具有正确的形式，即 $(\beta,0,\cdots,0)^{\mathrm{T}}$．设 A 原来的第 1 列用 A_1 表示．我们要求 $(I-vv^*)A_1=\beta e^{(1)}$，其中 $e^{(1)}$ 表示第一个标准单位向量 $e^{(1)}=(1,0,\cdots,0)^{\mathrm{T}}$．由 5.2 节引理 2 的证明，这个要求可以实现如下：先选择一个复数 β 使得 $|\beta|=\|A_1\|_2$ 且 $\langle A_1,\beta e^{(1)}\rangle$ 是实数．再取 $\alpha=\sqrt{2}/\|A_1-\beta e^{(1)}\|_2$，$v=\alpha(A_1-\beta e^{(1)})$．这样就允许 β 有两种选择，我们选择计算 v 的第一个分量时减法相消较少的那种．为帮助理解，我们把复数 β 和 a_{11} 写成极分解形式：

280

$$\beta = \|A_1\|_2 e^{i\varphi} \qquad a_{11} = |a_{11}| e^{i\theta}$$

于是，有

$$\langle A_1, \beta e^{(1)} \rangle = a_{11}\bar{\beta} = |a_{11}| \; \|A_1\|_2 e^{i(\theta-\varphi)}$$

这个值必须是实数，所以 $\theta-\varphi$ 应该是 0 或者是 π. 若我们选择 $\theta-\varphi=\pi$，则 v 的第一个分量将没有减法相消，因为

$$v_1 = \alpha(a_{11} - \beta) = \alpha(|a_{11}| e^{i\theta} - |\beta| e^{i(\theta-\pi)}) = \alpha(|a_{11}| + |\beta|) e^{i\theta}$$

于是，我们定义 β 为

$$\beta = -\|A_1\|_2 e^{i\theta} = -\|A_1\|_2 a_{11}/|a_{11}|$$

得到第一步中矩阵 U 的算法如下：

$$\beta \leftarrow -(a_{11}/|a_{11}|)\|A_1\|_2$$
$$y \leftarrow A_1 - \beta e^{(1)}$$
$$\alpha \leftarrow \sqrt{2}/\|y\|_2$$
$$v \leftarrow \alpha y$$
$$U \leftarrow I - vv^*$$

在 QR 分解中随后的步骤都类似于第一步. k 步之后，我们用 k 个酉阵左乘 A，结果得到一个前 k 列具有一定形式的矩阵；即前 k 列对角线下面都为 0. 这个情况可以总结为

$$U_k U_{k-1} \cdots U_1 A = \begin{bmatrix} J & H \\ 0 & W \end{bmatrix}$$

其中 J 是 $k\times k$ 上三角阵，0 是 $(m-k)\times k$ 零矩阵，H 是 $k\times(n-k)$ 矩阵，W 是 $(m-k)\times(n-k)$ 矩阵. 由前面的分析，存在向量 $v\in\mathbb{C}^{m-k}$ 使得 $I-vv^*$ 是 $m-k$ 阶酉阵并且 $(I-vv^*)W$ 的第 1 列第 1 个元素下面为 0，现在注意

$$\begin{bmatrix} I & 0 \\ 0 & I-vv^* \end{bmatrix}\begin{bmatrix} J & H \\ 0 & W \end{bmatrix} = \begin{bmatrix} J & H \\ 0 & (I-vv^*)W \end{bmatrix}$$

上式左边第一个因子是酉阵，我们用 U_{k+1} 来表示.

当 R 的第 $n-1$ 列取成适当的形式时，上述过程终止. 此时，我们有等式 $Q^*A=R$，其中 Q^* 表示所有那些用作因子的酉阵之积. 因为 Q 是酉阵，所以正如我们所希望的 $A=QR$. 这就是**豪斯霍尔德 QR 分解**. 等式 $Q^* = U_{n-1}U_{n-2}\cdots U_1$ 导致 $Q=U_1^*U_2^*\cdots U_{n-1}^*$. 从 U_k 的形式

281

$$U_k = \begin{bmatrix} I_{k-1} & 0 \\ 0 & I_{n-k+1} - vv^* \end{bmatrix}$$

可以看出 U_k 是埃尔米特的 $(U_k^* = U_k)$，因此

$$Q = U_1 U_2 \cdots U_{n-1}$$

例 1 我们用矩阵

$$A = \begin{bmatrix} 63 & 41 & -88 \\ 42 & 60 & 51 \\ 0 & -28 & 56 \\ 126 & 82 & -71 \end{bmatrix}$$

说明 QR 分解.

解 第一步，我们计算 β，因为 A_1 是实的，它可取为 $-\|A_1\|_2$：

$$\beta = -\|A_1\|_2 = -\|(63,42,0,126)^T\|_2 = -147$$

下面，我们计算 α：

$$\alpha = \sqrt{2}/\|A_1 - \beta e^{(1)}\|_2 = \sqrt{2}/\|(210,42,0,126)^T\|_2 = 1/(21\sqrt{70})$$

所以向量 v 由下式给出：

$$v = \alpha(A_1 - \beta e^{(1)}) = (10,2,0,6)^T/\sqrt{70}$$

于是第一个酉因子是

$$U_1 = I - vv^* = \frac{1}{35}\begin{bmatrix} -15 & -10 & 0 & -30 \\ -10 & 33 & 0 & -6 \\ 0 & 0 & 35 & 0 \\ -30 & -6 & 0 & 17 \end{bmatrix}$$

可算得积 U_1A 为

$$U_1A = \frac{1}{35}\begin{bmatrix} -5\,145 & -3\,675 & 2\,940 \\ 0 & 1\,078 & 2\,989 \\ 0 & -980 & 1\,960 \\ 0 & -196 & 1\,127 \end{bmatrix}$$

第二步相应的计算是

$$\beta = -\|(30.8, -28, -5.6)^T\|_2 = -42$$

$$\alpha = \sqrt{2}/\|(72.8, -28, -5.6)^T\|_2 = 0.018\,085$$

$$v = \alpha(1.316\,6, -0.506\,37, -0.101\,27)^T$$

$$I - vv^* = \begin{bmatrix} -0.733\,33 & 0.666\,67 & 0.133\,33 \\ 0.666\,67 & 0.743\,59 & -0.051\,28 \\ 0.133\,33 & -0.051\,28 & 0.989\,74 \end{bmatrix}$$

282

所以，第二个酉因子是

$$U_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -0.733\,33 & 0.666\,67 & 0.133\,33 \\ 0 & 0.666\,67 & 0.743\,59 & -0.051\,28 \\ 0 & 0.133\,33 & -0.051\,28 & 0.989\,74 \end{bmatrix}$$

并且我们有

$$U_2U_1A = \begin{bmatrix} -147 & -105 & 84 \\ 0 & -42 & -21 \\ 0 & 0 & 96.923\,1 \\ 0 & 0 & 40.384\,6 \end{bmatrix}$$

最后一步，类似的计算是

$$\beta = -\|(96.923\,1, 40.384\,6)^{\mathrm{T}}\|_2 = -105$$

$$\alpha = \sqrt{2}/\|(201.923\,1, 40.384\,6)^{\mathrm{T}}\|_2 = 0.006\,867\,7$$

$$v = (1.386\,75, 0.277\,35)^{\mathrm{T}}$$

$$I - vv^* = \begin{bmatrix} -0.923\,08 & -0.384\,62 \\ -0.384\,62 & 0.923\,08 \end{bmatrix}$$

所以，第三个因子是

$$U_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -0.923\,08 & -0.384\,62 \\ 0 & 0 & -0.384\,62 & -0.923\,08 \end{bmatrix}$$

因而，上三角阵 R 是

$$R = \begin{bmatrix} -147 & -105 & 84 \\ 0 & -42 & -21 \\ 0 & 0 & -105 \\ 0 & 0 & 0 \end{bmatrix} = 21\begin{bmatrix} -7 & -5 & 4 \\ 0 & -2 & -1 \\ 0 & 0 & -5 \\ 0 & 0 & 0 \end{bmatrix}$$

最后，我们有

$$Q^* = U_3U_2U_1 = \begin{bmatrix} -0.428\,57 & -0.285\,71 & 0 & -0.857\,14 \\ 0.095\,24 & -0.714\,29 & 0.666\,67 & 0.190\,48 \\ 0.476\,19 & -0.571\,43 & -0.666\,67 & -0.047\,62 \\ -0.761\,90 & -0.285\,71 & -0.333\,33 & 0.476\,19 \end{bmatrix}$$

$$= \frac{1}{21}\begin{bmatrix} -9 & -6 & 0 & -18 \\ 2 & -15 & 14 & 4 \\ 10 & -12 & -14 & -1 \\ -16 & -6 & -7 & 10 \end{bmatrix}$$

283

我们可以验证 $A=QR$；即

$$\begin{bmatrix} 63 & 41 & -88 \\ 42 & 60 & 51 \\ 0 & -28 & 56 \\ 126 & 82 & -71 \end{bmatrix} = \begin{bmatrix} -9 & 2 & 10 & -16 \\ -6 & -15 & -12 & -6 \\ 0 & 14 & -14 & -7 \\ -18 & 4 & -1 & 10 \end{bmatrix}\begin{bmatrix} -7 & -5 & 4 \\ 0 & -2 & -1 \\ 0 & 0 & -5 \\ 0 & 0 & 0 \end{bmatrix}$$

■

习题 5.3

1. 证明：若 $x \neq y$ 且 $\langle x, y\rangle$ 是实数，则由 $U=I-vu^*$ 给出的酉阵 U 满足 $Ux=y$，其中 $v=x-y$，$u=2v/\|v\|_2^2$。说明为什么这是构造豪斯霍尔德变换的一个较好的方法？假定 $\|x\|_2=\|y\|_2$.
2. 证明：若 A 是秩为 n 的 $m\times n$ 矩阵，则 A^*A 非奇异.
3. 证明：若 A 是秩为 n 的 $m\times n$ 矩阵，则 A^*A 是埃尔米特正定阵.
4. 设 $\{u_1, u_2, \cdots, u_n\}$ 是内积空间 X 的一个标准正交基. 证明：对 $x, y\in X$,

   a. $\|x\|_2^2 = \sum_{i=1}^{n} |\langle x, u_i\rangle|^2$

b. $\langle x,y\rangle = \sum_{i=1}^{n}\langle x,u_i\rangle\overline{\langle y,u_i\rangle}$

5. 设$\{u_1,u_2,\cdots,u_n\}$是内积空间 X 中子空间 U 的一个标准正交基.用等式 $Px=\sum_{i=1}^{n}\langle x,u_i\rangle u_i$ 定义 $P:X\to U$.证明

   a. P 是线性的.

   b. P 是幂等的($P^2=P$).

   c. 若 $x\in U$，则 $Px=x$.

   d. $\|Px\|_2\leqslant\|x\|_2$，对一切 $x\in X$.

   这个映射 P 称为 X 到 U 上的**正交投影**.

6. (续)设 $AB=Q$，其中 Q 是 $m\times n$，B 是 $n\times n$，$Q^*Q=I$. 证明 QQ^* 是$\mathbb{R}^m$ 到 A 的值域上的正交投影.

7. 酉阵的行列式是多少？正交矩阵的行列式是多少？

8. 正交集是线性无关的充要条件是什么？证明任何一个正交集都是线性无关的.

9. 对固定的 u 和 x，试问 t 的什么值可使表达式 $\|u-tx\|_2$ 达到极小？(在复数情况下有解.)

10. 若$\{x_1, x_2, \cdots, x_n\}$和$\{y_1, y_2, \cdots, y_n\}$是$\mathbb{C}^n$ 中的标准正交基，证明具有元素$\langle x_i, y_j\rangle$的矩阵是酉阵.

11. 使 $A^2=I$ 的矩阵 A 称为一个**对合**或者是**对合的**. 求 u 和 v 使 $I-uv^*$ 为一个对合的充要条件.

12. (续)设 $v^*v=2$，证明分块阵

$$\begin{bmatrix} I & 0 \\ 0 & I-vv^* \end{bmatrix}$$

   是一个对合.

13. 请问对合阵的乘积是对合的吗？

284

14. 若 U 和 V 是酉阵，那么下列矩阵

$$\begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix}$$

   是酉阵吗？

15. a. 证明所有阶数固定的酉阵集合是一个乘法群.

    b. 证明在运算 $A\to A^{\mathrm{T}}$，A^*，$\overline{A}$ 之下，上述集合是闭的.

16. 利用豪斯霍尔德算法求

$$\begin{bmatrix} 0 & -4 \\ 0 & 0 \\ -5 & -2 \end{bmatrix}$$

   的 QR 分解.

17. 证明：若 Q 是酉阵，则对一切 x 和 y，

$$\|x\|_2 = \|Qx\|_2 \text{和} \langle x,y\rangle = \langle Qx,Qy\rangle$$

   因此，在欧几里得空间中 Q(当其被看作一个变换时)保持长度、距离和角度. 利用从属于欧几里得范数的矩阵范数计算 $\|Q\|_2$.

18. 证明 A 是酉阵当且仅当其行形成一个标准正交系. (A 是一个方阵.)

19. 设 A 是 $m\times n$ 矩阵，b 是 m 维向量且 $\alpha>0$. 利用欧几里得范数，定义

$$F(x) = \|Ax-b\|_2^2+\alpha\|x\|_2^2$$

   证明：当 x 是下列方程

$$(A^{\mathrm{T}}A+\alpha I)x = A^{\mathrm{T}}b$$

的解时 $F(x)$ 达到极小. 并当如此定义 x 时，证明

$$F(x+h) = F(x) + (Ah)^T Ah + ah^T h$$

20. 设 D 是对角阵，U 是酉阵. 试问要对 D 作怎样进一步的假设我们才能推断 DU 是酉阵?
21. 当 U 和 AU 都是酉阵时，关于 A 可以得到怎样的结论?
22. 证明在求解方程 $Ax=b$ 的最小二乘问题中，我们可用 $CAx=Cb$ 替代正规方程，其中 C 是任何行-等价于 A^T 的 $n\times m$ 阵. 提示：回忆两个矩阵 G 和 H 都是行-等价的，如果存在一个非奇异矩阵 F 使得 $G=FH$.
23. 设 A 是秩为 n 的 $m\times n$ 矩阵. b 是 $\mathbb{R}^m$ 中的任意点. 证明集合

$$K_\lambda = \{x \in \mathbb{R}^n : \| Ax - b \|_2 \leqslant \lambda \}$$

是有界闭集(在 $\mathbb{R}^m$ 和 $\mathbb{R}^n$ 上的范数可以是任意的)
24. (续)假定假设条件如上题. 证明：若 $\lambda=2\|b\|_2$，则

285

$$\inf_{x\in \mathbb{R}^n} \| Ax - b \|_2 = \inf_{x\in K_\lambda} \| Ax - b \|_2$$

25. (续)若 A 是秩为 n 的 $m\times n$ 矩阵，则 $Ax=b$ 的最小二乘解满足不等式

$$\| x \|_2 \leqslant 2 \| b \|_2 \| B \|_2$$

其中 B 是 A 的任意一个左逆. 在 $\mathbb{R}^n$ 和 $\mathbb{R}^m$ 中使用欧几里得向量范数，$\|B\|_2$ 表示相应的从属矩阵范数.
26. 设 A 是一个未指定秩的 $m\times n$ 矩阵，$b\in\mathbb{R}^m$，又设

$$\rho = \inf\{ \| Ax - b \| : x \in \mathbb{R}^n \}$$

证明这个下确界可以达到. 换言之，即存在一个 x 使得 $\|Ax-b\|=\rho$. 本题中定义在 $\mathbb{R}^m$ 上的范数是任意的.
27. (续)沿用上题的假设，证明 $A^TAx=A^Tb$ 有解. 此处对 A 的秩不作任何假设.
28. 如果对 m 和 n 的相对值或 A 的秩不作任何假设. 证明对任何方程组 $Ax=b$，方程组 $A^TAx=A^Tb$ 肯定有解.
29. 给出满足 $\|x+y\|_2^2=\|x\|_2^2+\|y\|_2^2$ 和 $\langle x, y\rangle\neq 0$ 的向量 x 和 y 的例子.
30. 求下列方程组

$$[xy]\begin{bmatrix} 3 & 2 & 1 \\ 2 & 3 & 2 \end{bmatrix} = [3 \quad 0 \quad 1]$$

的最小二乘解.
31. (续)断言上题的解是 $x=29/21$ 和 $y=-2/3$. 如果不解 x 和 y，怎样来验证这个结论?
32. 设 A 是秩为 n 的 $(n+1)\times n$ 矩阵，z 是正交于 A 的列的非零向量. 证明方程 $Az+\lambda z=b$ 有一个 x 和 λ 的解. 并说明因此得到的 x 是方程 $Ax=b$ 的最小二乘解.
33. 下例是由 Noble and Daniel[1988]给出的，它指出在使用**不修正的格拉姆-施密特法**时舍入误差所产生的影响. 设 ε 是一个小的正数使得 $1+\varepsilon$ 和 $3+2\varepsilon$ 都是机器数但是 $3+2\varepsilon+\varepsilon^2$ 被计算成了 $3+2\varepsilon$. 现在对三个向量

$$v_1 = \text{fl}(1+\varepsilon,1,1)^T \quad v_2 = \text{fl}(1,1+\varepsilon,1)^T \quad v_3 = \text{fl}(1,1,1+\varepsilon)^T$$

应用格拉姆-施密特过程. 验证由计算机产生的标准正交基实际上是

$$x_1 = \text{fl}[1/\sqrt{3+2\varepsilon}]v_1 \quad x_2 = \text{fl}[1/\sqrt{2}](-1,1,0)^T \quad x_3 = \text{fl}[1/\sqrt{2}](-1,0,1)^T$$

注意：$\langle x_2, x_3\rangle=1/2$.
34. 试问对方程组 $Ax=b$ 作怎样的初等行运算能使所有的最小二乘解保持不变?
35. 设 A 是 $m\times n$ 矩阵且 $m>n$. A 有 QR 分解 $A=QR$. 设 Q' 是删去 Q 的后 $m-n$ 列所得到的 $m\times n$ 矩阵，R' 是删去 R 的后 $m-n$ 行得到的 $n\times n$ 矩阵. 证明 $A=Q'R'$.
36. 提供下列得到 $m\times n$ 矩阵 A 的 QR 分解方法的细节. 用 A_j、Q_j 和 R_j 分别表示 A、Q 和 R 的第 j 列. 证明 $A_j=$

$\sum_{k=1}^{j} r_{kj} Q_k$ 和 $r_{kj} = \langle A_j, Q_k \rangle$. 并利用这个等式决定 Q_1 和 r_{11}. 然后说明一旦得到结果 $Q_1, Q_2, \cdots, Q_{j-1}$ 和 $R_1, R_2, \cdots, R_{j-1}$ 后如何去求 Q_j 和 R_j.

286

37. 求矩阵 $\begin{bmatrix} 3 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$ 的 QR 分解.

38. 利用(10)式中的矩阵求 $\kappa_\infty(A^* A)$. 当 $\epsilon \to 0$ 时，会发生什么情况？

39. 对三个向量(3, 4, 0)、(1, 1, 1)和(1, 2, 0)依次应用格拉姆-施密特过程.

计算机习题 5.3

1. 同时编写格拉姆-施密特和修正格拉姆-施密特算法程序. 然后测试这两个程序看看哪个较好. 第一个测试涉及一个 20×10 矩阵，其元素是区间[0, 1]中均匀分布的随机数. 第二个测试也涉及一个 20×10 矩阵，但其元素是由某个初等函数生成的，例如

$$a_{ij} = \left(\frac{2i-21}{19}\right)^{j-1} \quad (1 \leqslant i \leqslant 20, 1 \leqslant j \leqslant 10)$$

在每一种情况下，从 A 生成 B，它的列都应该是标准正交的. 考察 $B^T B$ 看看它怎样靠近单位阵. 有关这种测试的进一步信息参见 Rice[1966].

2. 编写一个执行修正格拉姆-施密特算法的子程序或过程. 输入子程序的是秩为 n 的一个 $m \times n$ 矩阵 A，而输出的是矩阵 B 和 T.

3. (续)编写一个求方程组 $Ax=b$ 最小二乘解的子程序或过程. 这个过程应该调用上题中的过程.

4. 编写并测试豪斯霍尔德 QR 分解的程序.

5.4 奇异值分解和广义逆

奇异值分解是另外一种有着许多应用的矩阵分解. 我们首先研究一个描述其形式并断言其存在的定理.

定理 1(奇异值分解定理) 任意一个复 $m \times n$ 矩阵 A 可分解为

$$A = PDQ$$

其中 P 是 $m \times m$ 酉阵，D 是 $m \times n$ 对角阵，Q 是 $n \times n$ 酉阵.

证明 矩阵 $A^* A$ 是一个 $n \times n$ 埃尔米特阵. 因为

$$x^* (A^* A) x = (Ax)^* (Ax) \geqslant 0$$

287

所以它也是半正定的. 由此可得到 $A^* A$ 的特征值非负数(见习题 5.4.39)，现用 $\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2$ 来表示它们(在这个序列中，每个 σ_i^2 按其作为特征方程根的重数重复出现)，进一步，再把 σ_i 排序使 $\sigma_1^2, \sigma_2^2, \cdots, \sigma_r^2$ 都是正的，而 $\sigma_{r+1}^2, \sigma_{r+2}^2, \cdots, \sigma_n^2$ 都是 0. 设 $\{u_1, u_2, \cdots, u_n\}$ 是 $A^* A$ 的标准正交特征向量集，使得

$$A^* A u_i = \sigma_i^2 u_i$$

于是

$$\| A u_i \|_2^2 = u_i^* A^* A u_i = u_i^* \sigma_i^2 u_i = \sigma_i^2$$

这说明当 $i \geqslant r+1$ 时 $Au_i = 0$. 注意

$$r = \operatorname{rank}(A^* A) \leqslant \min\{\operatorname{rank}(A^*), \operatorname{rank}(A)\} \leqslant \min\{m, n\}$$

我们用 $u_1^*, u_2^*, \cdots, u_n^*$ 作为行构成一个 $n \times n$ 矩阵 Q. 接着，定义

$$v_i = \sigma_i^{-1} A u_i \qquad (1 \leqslant i \leqslant r)$$

对 $1 \leqslant i, j \leqslant r$，$v_i$ 构成一个标准正交系．我们有

$$v_i^* v_j = \sigma_i^{-1}(Au_i)^* \sigma_j^{-1}(Au_j) = (\sigma_i\sigma_j)^{-1}(u_i^* A^* A u_j) = (\sigma_i\sigma_j)^{-1}(u_i^* \sigma_j^2 u_j) = \delta_{ij}$$

我们选择额外的向量 v_i 使 $\{v_1, v_2, \cdots, v_m\}$ 是 $\mathbb{C}^m$ 的标准正交基．设 P 是 $m \times m$ 矩阵，其列是 $v_1, v_2, \cdots, v_m$．设 D 是 $m \times n$ 对角阵，$\sigma_1, \sigma_2, \cdots, \sigma_r$ 在其对角线上，其他地方都是0．于是

$$A = PDQ$$

为证明这一点，我们指出

$$P^* A Q^* = D$$

因为

$$(P^* A Q^*)_{ij} = v_i^* A u_j$$

所以，当 $j \geqslant r+1$ 时，上式为0，当 $j \leqslant r$ 时，上式是 $v_i^* \sigma_j v_j = \sigma_j \delta_{ij}$． ■

数 $\sigma_1, \sigma_2, \cdots, \sigma_n$（取非负数）称为 A 的**奇异值**．它们是 $A^* A$ 特征值的非负平方根．定理1中的分解 $A = PDQ$ 就是一个**奇异值分解**．

注意在证明的某些步中作了任意的选择．例如，$\sigma_1, \sigma_2, \cdots, \sigma_r$ 的次序是任意的，向量 $v_{r+1}, v_{r+2}, \cdots, v_m$ 也允许某种选择．因此，一个矩阵通常有许多个奇异值分解．

例1 求下列矩阵的一个奇异值分解

288

$$A = \begin{bmatrix} 7 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

解 我们有

$$A^* A = \begin{bmatrix} 49 & 0 & 0 & 0 \\ 0 & 9 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

因而，$\sigma_1 = 7$，$\sigma_2 = 3$，$\sigma_3 = 0$ 和 $\sigma_4 = 0$．由定理的证明，可以构成下面这些矩阵：

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad D = \begin{bmatrix} 7 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad Q = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

读者可以验证 A 的另一个奇异值分解是

$$\begin{bmatrix} 7 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 7 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

■

例2 求下列矩阵的一个奇异值分解

$$\begin{bmatrix} 0 & -1.6 & 0.6 \\ 0 & 1.2 & 0.8 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

解 再由定理可得

$$A^* A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

并取 $\sigma_1=1$，$\sigma_2=2$ 和 $\sigma_3=0$.（也可能是另一种次序.）选择特征向量并构成 Q，我们有(作为若干种选择之一)

$$Q = \begin{bmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

于是

$$v_1 = Au_1 = (0.6, 0.8, 0, 0)^*$$

$$v_2 = \frac{1}{2} Au_2 = (0.8, -0.6, 0, 0)^*$$

v_3 和 v_4 有好几种选择. 最简单的选择是

$$v_3 = (0,0,1,0)^* \quad v_4 = (0,0,0,1)^*$$

289

因而，一个奇异值分解如下：

$$\begin{bmatrix} 0 & -1.6 & 0.6 \\ 0 & 1.2 & 0.8 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.6 & 0.8 & 0 & 0 \\ 0.8 & -0.6 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

■

5.4.1 广义逆

对下列形式的一个 $m \times n$ 矩阵：

$$D = \begin{bmatrix} \sigma_1 & & & & & & \\ & \sigma_2 & & & & & \\ & & \ddots & & & & \\ & & & \sigma_r & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{bmatrix}$$

其中每个 σ_i 是正的，我们定义下列 $n \times m$ 矩阵为其**广义逆**：

$$D^+ = \begin{bmatrix} \sigma_1^{-1} & & & & & & \\ & \sigma_2^{-1} & & & & & \\ & & \ddots & & & & \\ & & & \sigma_r^{-1} & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{bmatrix}$$

在这两个矩阵中，未显示的元素都为0. 通过首先作 A 的一个奇异值分解

$$A = PDQ$$

再取

$$A^{+} = Q^{*} D^{+} P^{*}$$

来定义一个一般矩阵 A 的**广义逆**. 在后面我们将看到虽然一个矩阵的奇异值分解不唯一，但是其广义逆却是唯一确定的.

例3 例1中矩阵 A 的广义逆是什么?

解 因为此例中 P 和 Q 是单位阵，由此直接可得

290

$$A^{+} = \begin{bmatrix} 7^{-1} & 0 & 0 \\ 0 & 3^{-1} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

■

例4 求例2中矩阵 A 的广义逆.

解 例2中的结果给出

$$A^{+} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0.6 & 0.8 & 0 & 0 \\ 0.8 & -0.6 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & 0 & 0 \\ -0.4 & 0.3 & 0 & 0 \\ 0.6 & 0.8 & 0 & 0 \end{bmatrix}$$

■

5.4.2 不相容方程组和欠定方程组

广义逆的主要应用是不相容的或解不唯一的方程组. 考察写成下列矩阵形式的方程组

$$Ax = b$$

其中 A 是 $m \times n$ 矩阵，x 是 n 维向量，b 是 m 维向量. 这个问题的**最小解**定义如下：

1. 若方程组相容且有唯一解 x，则最小解定义为 x.

2. 若方程组相容且有一组解，则最小解是这组解中具有最小欧几里得范数的元素.

3. 若方程组不相容且有唯一的最小二乘解 x，则最小解定义为 x.

4. 若方程组不相容且有一组最小二乘解，则最小解是这组解中具有最小欧几里得范数的元素.

另一种定义叙述如下：利用欧几里得范数，设

$$\rho = \inf\{ \| Ax - b \|_2 : x \in \mathbb{C}^n \}$$

则方程 $Ax=b$ 的最小解是集合 $K = \{x : \| Ax - b \|_2 = \rho\}$ 中最小欧几里得范数的元素. 显然易见这个定义包含了前面所描述的四种情况. 例如，若 $\rho=0$，我们有情况1和2，而情况3和4则对应于 $\rho>0$.

定理2(广义逆最小解定理) 由广义逆给出的方程 $Ax=b$ 的最小解是

$$x = A^{+} b$$

[291]

证明 设 A 的一个奇异值分解是 $A=PDQ$. 设

$$c = P^{*} b \text{ 和 } y = Qx$$

当 x 取遍 $\mathbb{C}^n$ 时，因为 Q 是满射，即它把 $\mathbb{C}^n$ 映射到 $\mathbb{C}^n$ 上，所以 y 也取遍 $\mathbb{C}^n$. 因此，

$$\rho = \inf_x \| Ax - b \|_2 = \inf_x \| PDQx - b \|_2 = \inf_x \| P^{*}(PDQx - b) \|_2$$
$$= \inf_x \| DQx - P^{*} b \|_2 = \inf_y \| Dy - c \|_2$$

由矩阵 D 的特征，我们有

$$\| Dy - c \|_2^2 = \sum_{i=1}^{r} (\sigma_i y_i - c_i)^2 + \sum_{i=r+1}^{m} c_i^2$$

对 $1 \leqslant i \leqslant r$，取 $y_i = c_i/\sigma_i$，且允许 y_{r+1}，y_{r+2}，…，y_n 取任意值．那么上面的值最小．于是，我们有

$$\rho = \Big(\sum_{i=r+1}^{m} c_i^2 \Big)^{1/2}$$

在那些产生这个最小值 ρ 的所有 y 向量中，最小范数的向量有 $y_{r+1}=y_{r+2}=\cdots=y_n=0$. 这个向量是由下式给出的

$$y = D^{+} c$$

所以我们问题的最小解是

$$x = Q^{*} y = Q^{*} D^{+} c = Q^{*} D^{+} P^{*} b = A^{+} b$$

■

广义逆就像逆对可逆方程组那样，对不相容方程组或欠定方程组扮演着同样的角色．应该注意，任何方程 $Ax=b$ 的最小解都是唯一的，这是因为集合 K 是凸的并且有唯一的最小范数元素．

例 5 求下列方程组的最小解：

$$\begin{cases} 0x - 1.6y + 0.6z = 5 \\ 0x + 1.2y + 0.8z = 7 \\ 0x + 0y + 0z = 3 \\ 0x + 0y + 0z = -2 \end{cases}$$

解 其系数阵是例 2 中的 A. 而其广义逆在例 4 中已求得．因此，最小解是

$$A^{+} b = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -0.4 & 0.3 & 0 & 0 \\ 0.6 & 0.8 & 0 & 0 \end{bmatrix} \begin{bmatrix} 5 \\ 7 \\ 3 \\ -2 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 0.1 \\ 8.6 \end{bmatrix}$$

[292]

■

5.4.3 Penrose 性质

广义逆具有逆的某些(但不是一切)性质．例如，若 $n>m$，我们不能期望 $A^{+}A=I$ 成立．因为 A^{+}，A 和 $A^{+}A$ 的秩最多是 m 而 I 是 $n\times n$ 矩阵．然而，像 $AA^{+}A=A$ 那样的等式对任何 A 都成立．在一般情况下，下面这个定理涉及 4 个如此设定的等式．

定理 3(Penrose 性质定理) 对应于任何矩阵 A，至多存在一个具有下列 4 个性质的矩阵 X.

1. $AXA=A$.
2. $XAX=X$.
3. $(AX)^*=AX$.
4. $(XA)^*=XA$.

证明 设 X 和 Y 是两个具有性质 1-4 的矩阵. 那么通过系统地使用所指出的这些性质，我们有

$$
\begin{aligned}
&& \text{性质}\\
X &= XAX && 2\\
&= XAYAX && 1\\
&= XAYAYAYAX && 1\\
&= (XA)^*(YA)^*Y(AY)^*(AX)^* && 4 \text{ 和 } 3\\
&= A^*X^*A^*Y^*YY^*A^*X^*A^* &&\\
&= (AXA)^*Y^*YY^*(AXA)^* &&\\
&= A^*Y^*YY^*A^* && 1\\
&= (YA)^*Y(AY)^* &&\\
&= YAYAY && 4 \text{ 和 } 3\\
&= YAY && 2\\
&= Y && 2
\end{aligned}
$$

证毕.

这个定理是由 R. Penrose[1955]给出的，现在称条件 1-4 为 **Penrose 性质**. ■

定理 4(唯一广义逆定理) 矩阵的广义逆有 4 个 Penrose 性质. 因此，每个矩阵有唯一的广义逆.

证明 设 A 是任意的矩阵，其奇异值分解是

$$A = PDQ$$

则

293

$$A^+ = Q^* D^+ P^*$$

若 A 是 $m\times n$，则 D 也是 $m\times n$，并且有形式如下：

$$D_{ij} = \begin{cases}\sigma_i & 若\ i=j\leqslant r\\ 0 & 其他\end{cases}$$

由此可证

$$DD^+D=D$$

为此，我们记

$$(DD^+D)_{ij} = \sum_{v=1}^{n} D_{iv}\sum_{\mu=1}^{m} D^+_{v\mu}D_{\mu j}$$

除非 $i\leqslant r$ 和 $j\leqslant r$，因为存在 D_{iv} 和 $D_{\mu j}$，所以上式右边为 0. 于是，我们假定 $i\leqslant r$ 和 $j\leqslant r$，并

继续简化右边为

$$\sum_{v=1}^{r} D_{iv} \sum_{\mu=1}^{r} D_{v\mu}^{+} D_{\mu j} = \sigma_i \sum_{\mu=1}^{r} D_{i\mu}^{+} D_{\mu j} = \sigma_i \sigma_i^{-1} D_{ij} = D_{ij}$$

利用类似的方法，我们可证明 D^+ 有其余 3 个与 D 有关的 Penrose 性质．于是，证明关于 A^+ 的这 4 个性质就是一件很简单的事情了．例如，第一个性质可证明如下：

$$\begin{aligned} AA^+A &= PDQQ^*D^+P^*PDQ \\ &= PDD^+DQ \\ &= PDQ = A \end{aligned}$$ ■

其余性质的证明留作习题 5.4.28～29.

定理 5(奇异值分解性质定理)　如定理 1 证明中所述，设 A 有奇异值分解 $A=PDQ$，则我们有

1. A 的秩为 r.
2. $\{u_{r+1}, u_{r+2}, \cdots, u_n\}$ 是 A 的零空间的一个标准正交基.
3. $\{v_1, v_2, \cdots, v_r\}$ 是 A 的值域的一个标准正交基.
4. $\|A\|_2 = \max\limits_{1\leqslant i\leqslant n} |\sigma_i|$.

证明　因为 P 和 Q 非奇异，所以 A 的秩与 D 的秩相同，而 D 的秩显然是 r．若 $r<i\leqslant n$，则如定理 1 证明中指出的那样 $Au_i=0$．因为 A 的秩为 r，所以 A 的零空间的维数为 $n-r$．因此，$\{u_{r+1}, u_{r+2}, \cdots, u_n\}$ 是此零空间的一个基．又因为 A 的秩为 r，所以 A 的值域的维数为 r．如定理 1 证明中所指出的那样，$v_i=\sigma_i^{-1}Au_i$ 且 $\{v_1, v_2, \cdots, v_r\}$ 是矩阵 A 的值域的标准正交基．而 P 和 Q 又是酉阵，所以它们分别充当了$\mathbb{C}^m$ 和$\mathbb{C}^n$ 上的等距变换(保范映射)，因此　[294]

$$\begin{aligned} \|A\|_2 &= \sup\{\|Ax\|_2 : \|x\|_2 = 1\} \\ &= \sup\{\|PDQx\|_2 : \|x\|_2 = 1\} \\ &= \sup\{\|Dy\|_2 : \|y\|_2 = 1\} \\ &= \sup\{\sqrt{(\sigma_1 y_1)^2 + \cdots + (\sigma_n y_n)^2} : \|y\|_2 = 1\} \\ &= [\sup\{\sigma_1^2 y_1^2 + \cdots + \sigma_n^2 y_n^2 : \sum_{i=1}^{n} y_i^2 = 1\}]^{1/2} \\ &= [\max_{1\leqslant i\leqslant n} \sigma_i^2]^{1/2} = \max_{1\leqslant i\leqslant n} |\sigma_i| \end{aligned}$$ ■

下列定理称为**奇异值分解**的**紧凑形式**.

定理 6(奇异值分解：紧凑形式)　若 A 是一个秩为 r 的 $m\times n$ 矩阵，$m\geqslant n\geqslant r$，则 A 可以分解为 $A=VSU$，其中 V 是具有标准正交列的 $m\times r$ 矩阵，S 是一个非奇异的 $r\times r$ 对角阵，U 是具有标准正交行的 $r\times n$ 矩阵.

在定理 1 的证明中奇异值分解的几何解释是中内在的．在下列定理中给出一个更容易直观的解释.

定理 7(标准正交基定理)　设 L 是从$\mathbb{C}^m$ 到$\mathbb{C}^n$ 的一个线性变换，则存在$\mathbb{C}^m$ 的标准正交基 $\{u_1, u_2, \cdots, u_m\}$ 和$\mathbb{C}^n$ 的标准正交基 $\{v_1, v_2, \cdots, v_n\}$ 使得

$$Lu_i = \begin{cases} \sigma_i v_i & \text{若 } 1\leqslant i\leqslant \min(m,n) \\ 0 & \text{若 } \min(m,n) < i\leqslant m \end{cases}$$

证明 首先，设$\{e_1,e_2,\cdots,e_m\}$是$\mathbb{C}^m$的标准基，而$\{\bar{e}_1,\bar{e}_2,\cdots\bar{e}_n\}$是$\mathbb{C}^n$的标准基. 用等式$Le_i=\sum_{j=1}^{n}A_{ij}\bar{e}_j(1\leqslant i\leqslant m)$定义$m\times n$矩阵$A$. 选择$A$的一个奇异值分解$A=PDQ$. 用$u_1,u_2,\cdots,u_m$表示$P^*$的行，$v_1,v_2,\cdots,v_m$表示$Q^*$的列，则正如我们所见，$Lu_i=\sigma_i v_i$，其中$\sigma_i$是$m\times n$矩阵$D$的对角元. 下面就是证明这个结论的计算. 在$1,2,\cdots,m$范围中选定$k$，则

[295]

$$\begin{aligned}Lu_k&=L\left(\sum_{i=1}^{m}\langle u_k,e_i\rangle e_i\right)=\sum_{i=1}^{m}\langle u_k,e_i\rangle Le_i\\&=\sum_{i=1}^{m}(P^*)_{ki}\sum_{j=1}^{n}A_{ij}\bar{e}_j=\sum_{j=1}^{n}\sum_{i=1}^{m}(P^*)_{ki}A_{ij}\bar{e}_j\\&=\sum_{j=1}^{n}(P^*A)_{kj}\bar{e}_j=\sum_{j=1}^{n}(P^*A)_{kj}\sum_{s=1}^{n}\langle\bar{e}_j,v_s\rangle v_s\\&=\sum_{s=1}^{n}\sum_{j=1}^{n}(P^*A)_{kj}(Q^*)_{js}v_s=\sum_{s=1}^{n}(P^*AQ^*)_{ks}v_s\end{aligned}$$

等式$A=PDQ$导致$P^*AQ^*=D$. 矩阵D是一个对角元为A的奇异值σ_1，σ_2，$\cdots$，σ_ℓ的$m\times n$对角矩阵，$\ell=\min(m,\ n)$. 因此

$$Lu_k=\sum_{s=1}^{n}D_{ks}v_s=\begin{cases}\sigma_k v_k & 若\ k\leqslant\ell\\0 & 若\ k>\ell\end{cases}$$ ■

习题 5.4

1. 求下列这些矩阵的奇异值分解：

$$\begin{bmatrix}4&0&0\\0&0&0\\0&0&7\\0&0&0\end{bmatrix}\qquad\begin{bmatrix}2&1\end{bmatrix}\qquad\begin{bmatrix}5\\-4\end{bmatrix}$$

2. 求下列方程组的最小解：

a. $x_1+x_2=b_1$

b. $\begin{cases}x_1=b_1\\x_1=b_2\\x_1=b_3\end{cases}$

c. $\begin{cases}4x_1=b_1\\0x_1=b_2\\7x_3=b_3\\0x_2=b_4\end{cases}$

3. 当AA^*可逆时，求A^+.
4. 当$A^*A=I$时，求A^+.
5. 当A为埃尔米特和幂等时，即$A^*=A$和$A^2=A$时，求A^+.
6. 证明：若A是埃尔米特，则A^+也是埃尔米特.
7. 若A是埃尔米特正定阵，试问它的奇异值分解是什么？

8. 证明广义逆是矩阵的不连续函数. 提示：计算下列矩阵的广义逆.

$$\begin{bmatrix} 1 & 0 \\ 0 & \varepsilon \\ 0 & 0 \end{bmatrix}$$

9. 若 A 是埃尔米特矩阵，试问它的特征值与奇异值之间有什么关系？

296

10. 证明下列这些广义逆的性质：

    a. $A^{++}=A$

    b. $A^{*+}=A^{+*}$

11. 证明下列这些广义逆的性质：

    a. $(AA^*)^+=A^{+*}A^+$

    b. $A^+=A^*(AA^*)^+$

12. 用适当的例子来说明一般情况下 $(AB)^+\neq B^+A^+$.

13. 证明定理 6.

14. 参考定理 1 的证明来证明

$$A=\sum_{j=1}^{r}\sigma_j v_j u_j^*$$

15. (续)设 A 是秩为 r 的 $m\times n$ 矩阵. 如定理 1 中那样定义 u_i，v_i 和 σ_i. 证明

$$A^+=\sum_{j=1}^{r}\sigma_j^{-1}u_j v_j^*$$

    (与上题比较一下.)

16. (续)利用上面习题 5.4.14 的结果证明：当得到 A 的奇异值分解时，Ax 可用 $(n+m+1)r$ 次乘法和 $r(n+m-1)-m$ 次加法算出. 并把它与用 nm 次乘法和 $(n-1)m$ 次加法的直接乘法比较一下.

17. (续)如果把上面习题 5.4.14 中的求和在第 k 个被加数截断，我们能得到 A 的一个近似. 证明这样做的误差满足

$$\| A-\sum_{i=1}^{k}\sigma_i v_i u_i^* \|_2=\sigma_{k+1}$$

    其中使用的是从属于欧几里得向量范数的矩阵范数.

18. 设 A 是秩为 r 的 $m\times n$ 矩阵，$m\geqslant n\geqslant r$，奇异值分解 $A=PDQ$. 证明方程组 $Ax=b$ 相容当且仅当 $(P^*b)_i=0$，$r<i\leqslant m$.

19. 证明：若 A 是埃尔米特正定阵，则它的特征值和奇异值相同.

20. 证明：若两个矩阵酉等价，则它们的奇异值相同.（如果对适当的酉阵 U 和 V，有 $A=UBV$，就称 A 和 B 是酉等价的.）

21. 设 A 是具有奇异值 σ_1，σ_2，…，σ_n 的 $n\times n$ 矩阵. 证明 A 的行列式 $\det(A)=\pm\sigma_1\sigma_2\cdots\sigma_n$.

22. 设 $\|A\|_2$ 表示从属于欧几里得向量范数的矩阵范数. 并设 A 的奇异值是 $\sigma_1\geqslant\sigma_2\geqslant\cdots\geqslant\sigma_n$. 证明 $\|A\|_2=\sigma_1$.

23. 设 A 是具有奇异值分解 $A=PDQ$ 的方阵. 证明 A 的特征多项式是 $\pm\det(D-\lambda P^*Q^*)=0$.

24. 设 A 是一个 $m\times n$ 矩阵，X 是一个具有与 A 有关的 4 个 Penrose 性质的 $n\times m$ 矩阵. 证明方程组 $Ax=b$ 的最小解是 Xb.

25. 证明：若 A 是实的，则它有一个实的奇异值分解和一个实的广义逆.

297

26. 证明：元素为 $v_iu_i^*$ 的平方和为 1.（记号与定理 1 证明中的记号相同.）

27. 假如 $A=UDV$，其中 U 是 $m\times m$ 酉阵，V 是 $n\times n$ 酉阵，D 是 $m\times n$ 对角阵. 证明 $|d_{ii}|^2(1\leqslant i\leqslant n)$ 是 A^*A 的特征值.

28. 证明 D^+ 的其余 3 个 Penrose 性质(见定理 4).
29. (续)完成定理 4 的证明.
30. 证明：若 A 是秩为 n 的 $m\times n$ 矩阵，则 $A^+=(A^*A)^{-1}A^*$.
31. 证明：$m\times n$ 对角阵的广义逆是 $n\times m$ 对角阵.
32. 求任意 $m\times 1$ 矩阵和任意 $1\times n$ 矩阵的广义逆.
33. 证明$\mathbb{C}^m$ 到 A 的列空间上的正交投影是 AA^+.
34. 证明：若 A 对称，则 A^+ 也对称.
35. 证明：若 A 和 B 满秩，则$(AB)^+=B^+A^+$. 如果一个 $m\times n$ 矩阵 A 满足 $\mathrm{rank}(A)=\min(m, n)$，就称为它满秩的.
36. 求 uv^* 的广义逆，其中 u 和 v 是$\mathbb{C}^n$ 的元素.
37. 求一个元素全部为 1 的 $m\times n$ 矩阵的广义逆.
38. 利用定理 3 证明：若 B 是一个 $m\times r$ 矩阵，C 是一个 $r\times n$ 矩阵，并且 B, C 和 BC 的秩都为 r，则$(BC)^+=C^*(CC^*)^{-1}(B^*B)^{-1}B^*$.
39. 证明一个半正定矩阵的特征值都是非负的.

计算机习题 5.4

利用奇异值分解和广义逆，编写一个求方程组 $Ax=b$ 最小解的计算机程序.

5.5 特征值问题的弗朗西斯 *QR* 算法

在 5.2 节中，我们证明了舒尔定理. 按此定理任何方阵都酉相似于一个三角阵. 因而，可能有下面类型的分解式

$$UAU^*=T \tag{1}$$

其中 U 是酉阵，T 是三角阵. 因为 A 的特征值和 T 的相同，而三角阵的特征值就是它的对角元，所以我们将求出展示在 T 对角线上的 A 的特征值. 虽然我们知道对任何给定的矩阵 A，(1)式的分解存在，但计算它可不是一件简单的事情. 求 U 必定像求 n 次多项式的所有(复)根那样困难. 因为实际上我们是在计算 A 的特征多项式的根. 由习题 5.5.3 可知，每个多项式(除了一个纯量倍数外)都是一个矩阵的特征多项式.

298

5.5.1 *QR* 分解

弗朗西斯(Francis)的 *QR* 算法[1961]是一个通过设计产生(1)式中的 T 来求出 A 的特征值的迭代法. 从它的名字可以推断出这个算法利用了 *QR* 分解. 其中，所有的矩阵为 $n\times n$ 方阵.

在 5.3 节中讨论了产生下列分解的算法

$$A=QR \tag{2}$$

其中 Q 是酉阵，R 是上三角阵. 这里，我们需要稍微改进一下这个分解式. 即我们希望 R 有非负的对角元. 这是很容易做到的. 事实上，若已知(2)式而 R 没有非负的对角元，我们即可定义某个对角酉阵 D 并由下式代替(2)式

$$A=(QD)(D^*R)=\tilde{Q}\tilde{R} \tag{3}$$

$D=\mathrm{diag}(d_{ii})$的定义是当 $r_{ii}\neq 0$ 时 $d_{ii}=r_{ii}/|r_{ii}|$，而当 $r_{ii}=0$ 时 $d_{ii}=1$. 很容易验证 D 是酉阵且矩阵 $\tilde{R}=D^*R$ 有非负的对角元.

QR 算法的基本形式如下：

```
A₁ ← A
for k=1 to M do
    QR 分解：A_k = Q_k R_k，其中 Q_k 是酉阵而 R_k 是有非负对角元的上三角阵
    A_{k+1} ← R_k Q_k
end do
```

如果情况良好的话，A_k 的对角元收敛(当 $k\to\infty$)于一个向量，其分量为 A 的特征值.

实际上，基本的 QR 算法组合了好几个附加的过程以节约运算和加快收敛. 这些内容不久将加以讨论. 首先，我们注意到在算法中产生的所有矩阵 A_k 是酉相似于 A 的. 这可从下列等式看出

$$A_k = Q_kR_k = (Q_kR_k)(Q_kQ_k^*) = Q_kA_{k+1}Q_k^*$$

其次，我们注意到，若 A 是实的，则随后的 A_k 也是实的. 因此，如果 A 有某些非实特征值，那么我们只能期望，在最好的情况下，A_k 将收敛于一个其对角线上为 2×2 子阵的"三角形"矩阵.

5.5.2 约化到上海森伯格形

为了节约 QR 迭代中所涉及的运算量，首先利用酉相似变换把矩阵 A 约化到**上海森伯格形**. 上海森伯格矩阵是这样的一个矩阵 H：当 $i>j+1$ 时，其元 $h_{ij}=0$. 因此，H 具有下列形式 [299]

$$H=\begin{bmatrix} * & * & * & * & \cdots & * & * & * & * \\ * & * & * & * & \cdots & * & * & * & * \\ 0 & * & * & * & \cdots & * & * & * & * \\ 0 & 0 & * & * & \cdots & * & * & * & * \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ddots & * & * & * & * \\ 0 & 0 & 0 & 0 & \cdots & * & * & * & * \\ 0 & 0 & 0 & 0 & \cdots & 0 & * & * & * \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & * & * \end{bmatrix}$$

使用豪斯霍尔德算法的酉相似变换约化 A 为 H. 第一步，把第 1 列约化到适当的形式，第二步，把第 2 列约化到适当的形式，如此等等. 下面描述第 k 步. 在第 k 步开始时，第 1 列到第 $k-1$ 列已经有了上海森伯格矩阵的适当形式. 我们把部分约化的矩阵按下列方式分块，其中维数已被指明：

$$\begin{bmatrix} B_{k\times k} & C \\ D & E_{(n-k)\times(n-k)} \end{bmatrix}$$

式中的 B 是一个 $k\times k$ 上海森伯格阵. D 是 $(n-k)\times k$ 矩阵，除了第 k 列外其余均为 0. C 是 $k\times(n-k)$ 矩阵，而 E 是 $(n-k)\times(n-k)$ 矩阵. 最后两个矩阵 C 和 E 没有什么特殊的结构. 设 U 是任意的 $n-k$ 阶酉阵，则

$$\begin{bmatrix} I & 0 \\ 0 & U \end{bmatrix}\begin{bmatrix} B & C \\ D & E \end{bmatrix}\begin{bmatrix} I & 0 \\ 0 & U^* \end{bmatrix}=\begin{bmatrix} B & CU^* \\ UD & UEU^* \end{bmatrix} \tag{4}$$

我们将选择 U 使得 UD 的第 k 列是向量 $(\beta, 0, 0, \cdots, 0)^T$. 注意 D 是下列形式的矩阵

$$D=\begin{bmatrix}0 & \cdots & 0 & d_1\\ 0 & \cdots & 0 & d_2\\ \vdots & \ddots & \vdots & \vdots\\ 0 & \cdots & 0 & d_{n-k}\end{bmatrix}$$

所以要确定 U 使得

$$Ud=\beta e^{(1)}\text{，其中 } d=\begin{bmatrix}d_1\\ d_2\\ \vdots\\ d_{n-k}\end{bmatrix}$$

就足够了．其中 $e^{(1)}$ 表示 $\mathbb{C}^{n-k}$ 中的第一个标准单位向量．在积 UD 中，前 $k-1$ 列自动为 0，而第 k 列在 β 下面的元素都为 0．

300

由 5.2 节中的引理 2，我们看到应该选择 β 使得 $\langle d,\beta e^{(1)}\rangle$ 是实数且 $\|\beta e^{(1)}\|_2=\|d\|_2$．我们取

$$U=I-vv^*\text{，其中 } v=\alpha(d-\beta e^{(1)}) \tag{5}$$

已选取 $\beta=-(d_1/|d_1|)\|d\|_2$ 和 $\alpha=\sqrt{2}/\|d-\beta e^{(1)}\|_2$．（这些细节来自引理 2 的证明和 5.3 节中豪斯霍尔德分解的证明.）

例 1 利用酉相似变换约化下列矩阵为上海森伯格形式.

$$A=\begin{bmatrix}[1] & [2\quad 3\quad 4]\\ \begin{bmatrix}4\\2\\4\end{bmatrix} & \begin{bmatrix}5&6&7\\1&5&0\\2&1&0\end{bmatrix}\end{bmatrix}$$

解 注意，我们已经按前面列出过程中的第 1 步所需要的方式把 A 分块．开始我们取 $d=(4,2,4)^{T}$．因为第一个分量是实的，所以就假设 $\beta=-\|d\|_2=-6$ 且 $\alpha=1/\sqrt{60}$．利用公式(5)．我们得到

$$v=\frac{1}{\sqrt{60}}(10,2,4)^{T}$$

因而第一个 U 阵是

$$U=I-vv^*=\frac{1}{15}\begin{bmatrix}-10 & -5 & -10\\ -5 & 14 & -2\\ -10 & -2 & 11\end{bmatrix}$$

执行(4)式中指出的乘法完成了第一步．结果是

$$UAU^*=\begin{bmatrix}\begin{bmatrix}1 & -5\\ -6 & \frac{385}{45}\end{bmatrix} & \begin{bmatrix}\frac{72}{45} & \frac{54}{45}\\ -\frac{163}{45} & \frac{34}{45}\end{bmatrix}\\ \begin{bmatrix}0 & \frac{62}{45}\\ 0 & \frac{259}{45}\end{bmatrix} & \begin{bmatrix}\frac{677}{152} & -\frac{311}{225}\\ -\frac{536}{225} & -\frac{352}{225}\end{bmatrix}\end{bmatrix}$$

$$
= \begin{bmatrix} \begin{bmatrix} 1 & -5 \\ -6 & 8.5556 \end{bmatrix} & \begin{bmatrix} 1.6 & 1.2 \\ -3.6222 & 0.75556 \end{bmatrix} \\ \begin{bmatrix} 0 & 1.3778 \\ 0 & 5.7556 \end{bmatrix} & \begin{bmatrix} 3.0089 & -1.3822 \\ -2.3822 & -1.5644 \end{bmatrix} \end{bmatrix}
$$

在第二步中，把当前的部分约化矩阵按前面等式中所指出的那样分块. 计算导致 [301]

$$
\begin{aligned}
d &= (1.3778,\quad 5.7555)^{T} \\
\beta &= -5.9182 \\
\alpha &= 0.15218 \\
v &= (1.1103, 0.87590)^{T} \\
U &= \begin{bmatrix} -0.23280 & -0.97252 \\ -0.97252 & 0.23280 \end{bmatrix}
\end{aligned}
$$

再次执行(4)式，就得到最终的上海森伯格矩阵

$$
H = \begin{bmatrix} 1 & -5 & -1.5395 & -1.2767 \\ -6 & 8.5556 & 0.10848 & 3.6986 \\ 0 & -5.9182 & -2.1689 & -1.1428 \\ 0 & 0 & -0.14276 & 3.6133 \end{bmatrix}
$$

■

虽然计算是在一台像 Marc-32 那样具有 7 位有效数字的计算机上执行的，但是这里显示的是舍入后的结果.

5.5.3 位移 QR 分解

下面的方法是把基本 QR 算法与反复的**原点位移**结合起来. 在讨论这个方法之前，我们要指出为什么它是必要的.

例 2 我们把基本 QR 算法应用于例 1 中 4×4 矩阵所产生的结果展示出来. 因而，用基本的迭代

$$
A_k = Q_k R_k, A_{k+1} = R_k Q_k
$$

生成一系列矩阵 A_k.

解 展示的结果已被舍入到 5 位有效数字.

$$
A_1 \leftarrow \text{Hessenberg}(A)
$$

$$
A_2 = \begin{bmatrix} 10.135 & 1.9821 & -0.75082 & 5.5290 \\ 6.7949 & -2.8402 & 0.52664 & 1.2616 \\ 0 & 0.19692 & 1.5057 & 1.7031 \\ 0 & 0 & 1.7508 & 2.1994 \end{bmatrix}
$$

$$\vdots$$

$$
A_{10} = \begin{bmatrix} 11.105 & -4.7599 & 3.8826 & -4.0296 \\ -0.00045570 & -3.8487 & -0.72647 & 1.2553 \\ 0 & -0.068658 & 3.5669 & 0.16324 \\ 0 & 0 & 0 & 0.17645 \end{bmatrix}
$$

$$\vdots$$

[302]

$$A_{20}=\begin{bmatrix}11.106 & -4.7403 & 3.9060 & -4.0296\\ 0 & -3.8526 & -0.68985 & 1.2559\\ 0 & -0.032156 & 3.5706 & 0.15706\\ 0 & 0 & 0 & 0.17645\end{bmatrix}$$

我们注意到所希望的上三角形式并未迅速地实现，因为(3，2)位置中的元素远离 0. 在此例中缓慢地收敛于一个上三角阵显然是相当棘手的，尽管一个特征值 0.176 45 在第 10 步已被非常准确地确定了. 另一个特征值 11.106 也很快地确定了，但中间的其他两个特征值在第 20 步仅能估计出 2 位或 3 位有效数字. ■

基本算法缓慢的收敛性可以通过对随后的矩阵执行位移来缓解，其中位移定义为用 $A-zI$ 来替代矩阵 A. **位移 QR 算法**以下列方式进行：

```
A_1 ← Hessenberg(A)
for k=1 to M do
    给出 A_k，计算纯量 z_k 同时说明下面的 QR 分解：A_k - z_k I = Q_k R_k
    A_{k+1} ← R_k Q_k + z_k I
end do
```

若算法中的纯量 z_k 取 A_k 右下方的对角元，则迭代将迅速地在最后一行产生一个形如$(0, 0, \cdots, 0, \alpha)^{\mathrm{T}}$ 的向量. 因而数 α 是 A 的一个特征值. 其后最佳的方法是利用 5.2 节中说明过的方式划去最后的行和列，以降低矩阵的阶数并重复这整个过程使得矩阵的阶数越来越小. 为减轻计算的负担，对大的矩阵先初始约化到海森伯格形是可取的.

引理 1(谱引理) 设矩阵 A 分块成如下形式：

$$A=\begin{bmatrix}B & C\\ 0 & E\end{bmatrix}$$

其中 B 和 E 是方阵，则 A 的谱(即它的特征值集)是 B 和 E 谱的并.

证明 等式 $Ax=\lambda x$ 的分块形式是

$$\begin{bmatrix}B & C\\ 0 & E\end{bmatrix}\begin{bmatrix}u\\ v\end{bmatrix}=\lambda\begin{bmatrix}u\\ v\end{bmatrix} \tag{6}$$

303 或等价地

$$\begin{cases}Bu+Cv=\lambda u\\ \qquad\ Ev=\lambda v\end{cases}$$

若 λ 是 A 的特征值，则(6)式有非平凡解$(u, v)^{\mathrm{T}}$. 若 $v\neq 0$，则 λ 是 E 的特征值. 若 $v=0$，则 $u\neq 0$，且 λ 是 B 的特征值. 这就证明了 $\mathrm{sp}(A)\subseteq\mathrm{sp}(B)\cup\mathrm{sp}(E)$.

反之，若 λ 是 B 的一个特征值，而 u 是相应的(非零)特征向量，则$(u, 0)^{\mathrm{T}}$ 是(6)式的解. 若 λ 是 E 的一个特征值但不是 B 的特征值，则设 v 是满足 $Ev=\lambda v$ 的非零向量. 下面，解方程$(B-\lambda I)u=-Cv$. 因为 λ 不是 B 的特征值，所以这个方程可解. 因而向量$(u, v)^{\mathrm{T}}$ 是(6)式的解. 这就证明了 $\mathrm{sp}(B)\cup\mathrm{sp}(E)\subseteq\mathrm{sp}(A)$. ■

例 3 对例 1 的海森伯格阵 H 应用位移 QR 算法.

解 下面给出 5 次位移 QR 分解和矩阵收缩的结果.

$$A\rightarrow H$$

$$H \to A_5 = \begin{bmatrix} 2.6141 & -10.087 & -2.4480 & -2.4727 \\ -5.5345 & 4.6668 & 3.5719 & 2.8753 \\ 0 & -0.28730 & 0.14546 & 0.10900 \\ 0 & 0 & 0 & 3.5736 \end{bmatrix}$$

$$\text{deflate}(A_5) \to \tilde{a}_5 = \begin{bmatrix} 11.001 & -5.0329 & -4.1730 \\ -0.30955 & -3.7507 & 1.3719 \\ 0 & 0 & 0.17645 \end{bmatrix}$$

$$\text{deflate}(\tilde{a}_5) \to \hat{A}_5 = \begin{bmatrix} 11.106 & -4.7234 \\ 0 & -3.8556 \end{bmatrix}$$

显然，计算的特征值是 3.573 6，0.176 45，11.106 和－3.855 6．它们精确到所示的小数位数． ■

5.5.4 初等行运算和列运算

可以用另一种(更为简单的)方法来约化一个矩阵为海森伯格形．这个方法使用初等行运算和列运算．每一步都用矩阵 E_i 左乘并用一个矩阵 E_i^{-1} 右乘当前的矩阵；因而 $A \leftarrow E_i A E_i^{-1}$．逆 E_i^{-1} 容易确定(见习题 4.1.5)．行和列运算一次只执行一种，从左到右进行．例 4 指出了这项工作是如何完成的．

例 4 我们打算用相似变换约化下列矩阵为上海森伯格形． [304]

$$A = \begin{bmatrix} -3 & 3 & 7 & 2 \\ 1 & 2 & 3 & -5 \\ 2 & -1 & 0 & 3 \\ 4 & 2 & -2 & 4 \end{bmatrix}$$

解 首先，如果我们不选主元，可从第 3 行和第 4 行中减去第 2 行的倍数产生所要求的 0．然后，像在高斯消元法中那样，我们要求乘数是小的．因此，交换第 2 行和第 4 行，把较强的主元素放在恰当的位置．(这里，我们仅使用行主元而不是尺度的行主元．)于是，为了有一个相似变换交换第 2 列和第 4 列是必要的．(为简化这个过程，此处我们实际上在做交换而不像 4.3 节高斯消元法中那样使用一个指标数组．)所得的矩阵是

$$A \to \begin{bmatrix} -3 & 3 & 7 & 2 \\ 4 & 2 & -2 & 4 \\ 2 & -1 & 0 & 3 \\ 1 & 2 & 3 & -5 \end{bmatrix} \to \begin{bmatrix} -3 & 2 & 7 & 3 \\ 4 & 4 & -2 & 2 \\ 2 & 3 & 0 & -1 \\ 1 & -5 & 3 & 2 \end{bmatrix}$$

下面，第 3 行减去第 2 行的 1/2，第 4 行减去第 2 行的 1/4．接着，再作逆的列运算；即把第 3 列的 1/2 加到第 2 列，第 4 列的 1/4 加到第 2 列．(见习题 4.1.3.)结果是

$$\begin{bmatrix} -3 & 2 & 7 & 3 \\ 4 & 4 & -2 & 2 \\ 0 & 1 & 1 & -2 \\ 0 & -6 & \frac{7}{2} & \frac{3}{2} \end{bmatrix} \to \begin{bmatrix} -3 & \frac{11}{2} & 7 & 3 \\ 4 & 3 & -2 & 2 \\ 0 & \frac{3}{2} & 1 & -2 \\ 0 & -\frac{17}{4} & \frac{7}{2} & \frac{3}{2} \end{bmatrix} \to \begin{bmatrix} -3 & \frac{25}{4} & 7 & 3 \\ 4 & \frac{7}{2} & -2 & 2 \\ 0 & 1 & 1 & -2 \\ 0 & -\frac{31}{8} & \frac{7}{2} & \frac{3}{2} \end{bmatrix}$$

在处理第2列时，我们有必要再次求主元．通过执行适当的行运算和逆的列运算使我们得到了下列3个矩阵，其中最后面的那个矩阵就是我们所要求的．

$$\begin{bmatrix} -3 & \frac{25}{4} & 7 & 3 \\ 4 & \frac{7}{2} & -2 & 2 \\ 0 & -\frac{31}{8} & \frac{7}{2} & \frac{3}{2} \\ 0 & 1 & 1 & -2 \end{bmatrix} \to \begin{bmatrix} -3 & \frac{25}{4} & 3 & 7 \\ 4 & \frac{7}{2} & 2 & -2 \\ 0 & -\frac{31}{8} & \frac{3}{2} & \frac{7}{2} \\ 0 & 1 & -2 & 1 \end{bmatrix} \to \begin{bmatrix} -3 & \frac{25}{4} & -\frac{37}{8} & 3 \\ 4 & \frac{7}{2} & -\frac{39}{4} & 2 \\ 0 & -\frac{31}{8} & \frac{3}{2} & \frac{7}{2} \\ 0 & 0 & -\frac{50}{31} & \frac{59}{31} \end{bmatrix}$$ ■

习题 5.5

1. 设 A 是一个 $n\times n$ 上海森伯格阵，在位置 $A_{k,k-1}$ 上为0．证明 A 的谱是两个子阵 $A_{i,j}(1\leqslant i,\ j<k)$ 和 $A_{i,j}(k<i,\ j\leqslant n)$ 谱的并．

[305]

2. 证明在 QR 算法中，我们有 $A_{k+1}=Q_k^*A_kQ_k$．并由此证明 A^k 的 QR 分解是 $(Q_1Q_2\cdots Q_k)(R_kR_{k-1}\cdots R_1)$．

3. 设 $a_0,\ a_1,\ \cdots,\ a_{n-1}$ 是任意的复数．取

$$p(t)=a_0+a_1t+a_2t^2+\cdots+a_{n-1}t^{n-1}+t^n$$

定义这个多项式的**友阵**为 $n\times n$ 矩阵

$$A=\begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ddots & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \\ -a_0 & -a_1 & -a_2 & -a_3 & \cdots & -a_{n-2} & -a_{n-1} \end{bmatrix}$$

证明：$(-1)^np$ 是 A 的特征多项式．提示：利用归纳法证明 $\det(A-\lambda I)=(-1)^np_n(\lambda)$．为了约化 $A-\lambda I$ 的行列式，用它的第 n 列元素展开它．

4. （续）对上题中的矩阵 A，证明

$$A^n=-\sum_{i=0}^{n-1}a_iA^i$$

提示：可以直接证明，或者利用上题和凯莱-哈密顿定理．

5. 证明对任何 n 个整数的集合，存在一个 $n\times n$ 整数矩阵，它的谱是给定整数的集合．

6. 求下列矩阵的特征值．

$$\begin{bmatrix} -1 & -4 & 1 \\ -1 & -2 & -5 \\ 5 & 4 & 3 \end{bmatrix}$$

7. 证明在位移 QR 算法中，A_{k+1} 酉相似于 A_k．

8. 求方程 $Ax=\lambda Bx$ 非平凡解 λ 值的问题是**广义特征值问题**．说明当 B 非奇异时这个问题可改写为通常的特征值问题．

9. （续）说明：在我们求问题 $Ax=\mu(B+tA)x$ 中的广义特征值时，倘若 $\mu t\neq1$，则很容易就算出 $Ax=\lambda Bx$ 的特

征值.

10. 选择整数 p 和 q 使得 $1\leqslant p<q\leqslant n$. 选择复数 α 和 β 使得 $|\alpha|^2+|\beta|^2=1$. 设 U 除了 4 个元素：$U_{pp}=U_{qq}=\alpha$ 和 $U_{pq}=-U_{qp}=\beta$ 之外，是 $n\times n$ 单位矩阵. 假定 $\alpha\bar{\beta}$ 是实数，证明 U 是酉阵.

11. 设 A 是有下列上三角块结构的实阵

$$A=\begin{bmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1n} \\ 0 & A_{22} & A_{23} & \cdots & A_{2n} \\ 0 & 0 & A_{33} & \cdots & A_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & A_{nn} \end{bmatrix}$$

306

其中 A_{ii} 是 2×2 矩阵. 给出计算 A 的特征值的一个简单过程，包括证明.

12.（续）参考上题，证明 A 奇异当且仅当至少有一个 A_{ii} 奇异.

13. 证明或否定：若 U 是酉阵，R 是上三角阵且 UR 是上海森伯格阵，则 U 是上海森伯格阵.

14. 证明：若 T 是上三角可逆阵，则 T^{-1} 也是上三角可逆阵.

15. 证明：若 A 是上海森伯格阵，T 是上三角阵，则 AT 和 TA 是上海森伯格阵.

16. 证明：若 T 是上三角阵，AT 是上海森伯格阵，则 TA 是上海森伯格阵.（假设 A 和 T 都是 $n\times n$ 矩阵，但不一定可逆.）

17.（续）如上题和例 4 中略述的在约化一个 $n\times n$ 矩阵为上海森伯格形中大约需要多少次乘法？

计算机习题 5.5

1. 根据(4)式编写约化一个矩阵为海森伯格形的计算机子程序或过程并应用于例 1 中的矩阵.

2.（续）增加一个执行基本 QR 算法的计算机子程序或过程并再次产生例 2 的结果.

3.（续）修改计算机子程序或过程使之执行位移 QR 算法. 再产生例 3 的结果. 此外利用你的程序求下列矩阵的特征值（343，294 和 147±196i）：

$$\begin{bmatrix} 190 & 66 & -84 & 30 \\ 66 & 303 & 42 & -36 \\ 336 & -168 & 147 & -112 \\ 30 & -36 & 28 & 291 \end{bmatrix}$$

4. 利用例 4 中那样的相似变换序列编写约化一个矩阵为上海森伯格形的过程或子程序. 每个相似变换应该跟随此例中的算法，即逆初等列运算跟在初等行运算后面. 对例 4 的矩阵测试这个程序. 对例 1 中的矩阵应用它并把它与课本中给出的过程相比较.

307

第6章 函数逼近

6.0 概述

在这一章中，讨论在计算机内函数的表示问题．我们将考虑几个不同的子问题．无论已知相对较少的点还是相对较多的点(或者全部点)，这种函数表示根据函数被表示的类型不同而不同．所选择的函数表示(多项式、样条函数、连分式等)也决定其理论特性．我们从最古老且最简单的多项式表示开始．

6.1 多项式插值

在这一节中，我们解决下述问题：给定一个有 $n+1$ 个数据点 (x_i, y_i) 的数表：

x	x_0	x_1	x_2	$\cdots$	x_n
y	y_0	y_1	y_2	$\cdots$	y_n

我们找一个次数尽可能低的多项式 p，使得

$$p(x_i) = y_i \qquad (0 \leqslant i \leqslant n)$$

308

这样的多项式称为这组数据点上的**插值**多项式．下面是解答这个问题的定理．

定理 1(多项式插值定理) *若 x_0，x_1，$\cdots$，x_n 是不同的实数，则对任意数值 y_0，y_1，$\cdots$，y_n，存在唯一的次数至多是 n 次的多项式 p_n 使得*

$$p_n(x_i) = y_i \qquad (0 \leqslant i \leqslant n)$$

证明 先证明其唯一性．假设有两个这样的多项式 p_n 和 q_n．那么，对于 $0 \leqslant i \leqslant n$，多项式 $p_n - q_n$ 具有性质 $(p_n - q_n)(x_i) = 0$．由于 $p_n - q_n$ 最多是 n 次多项式，如果它不是零多项式，那么它最多有 n 个零点．因为 x_i 是互不相同的，因而 $p_n - q_n$ 有 $n+1$ 个零点，所以它一定是零多项式．因此，$p_n \equiv q_n$．

对于定理的存在性部分，我们用数学归纳法证明．当 $n=0$ 时，可以选择一个常值函数 p_0 (多项式的次数 $\leqslant 0$)，使得 $p_0(x_0) = y_0$，假设已经得到一个次数不超过 $k-1$ 次的多项式 p_{k-1} 使得 $p_{k-1}(x_i) = y_i$，$0 \leqslant i \leqslant k-1$．我们尝试构造如下形式的 p_k

$$p_k(x) = p_{k-1}(x) + c(x - x_0)(x - x_1)\cdots(x - x_{k-1}) \tag{1}$$

毫无疑问，这是一个次数至多是 k 次的多项式．因为

$$p_k(x_i) = p_{k-1}(x_i) = y_i \qquad (0 \leqslant i \leqslant k-1)$$

所以 p_k 插值了 p_{k-1} 插值的那些数据．现在我们根据条件 $p_k(x_k) = y_k$ 来确定未知系数 c．由这个条件可以得到方程

$$p_{k-1}(x_k) + c(x_k - x_0)(x_k - x_1)\cdots(x_k - x_{k-1}) = y_k \tag{2}$$

因为与 c 相乘的因子非零(为什么?)，所以从(2)式无疑可以求出 c． ■

6.1.1 牛顿型插值多项式

我们要给出一个执行上述证明中递归过程的算法．在此之前，我们先作一些观察．首先，上述证明中多项式 p_0，p_1，$\cdots$，p_n 中的每一个 p_k 可以由 p_{k-1} 添加一个简单项而得到．因此，

在递归过程的最后，p_n 就是一些项的和，而且 p_0，p_1，…，p_{n-1}可以清晰地反映在 p_n 的表达式中. 每个多项式 p_k 具有形式

309

$$p_k(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + \cdots + c_k(x - x_0)\cdots(x - x_{k-1}) \tag{3}$$

它的紧凑形式是

$$p_k(x) = \sum_{i=0}^{k} c_i \prod_{j=0}^{i-1}(x - x_j) \tag{4}$$

当 $m<0$ 时，我们已经约定 $\prod_{j=0}^{m}(x - x_j) = 1$. (4)式的最初几项是

$$p_0(x) = c_0$$

$$p_1(x) = c_0 + c_1(x - x_0)$$

$$p_2(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1)$$

依此类推. 这些多项式称为**牛顿型插值多项式**.

假设已知系数 c_0，c_1，…，c_k，为了计算 p_k，要用到一个有效的方法，称为**嵌套乘法**或者**霍纳算法**. 这一点很容易由下面形式的一个任意表达式得到说明

$$u = \sum_{i=0}^{k} c_i \prod_{j=0}^{i-1} d_j = c_0 + c_1 d_0 + c_2 d_0 d_1 + \cdots + c_k d_0 d_1 \cdots d_{k-1} \tag{5}$$

其想法是把上式写成下列形式：

$$u = (\cdots(((c_k)d_{k-1} + c_{k-1})d_{k-2} + c_{k-2})d_{k-3} + \cdots + c_1)d_0 + c_0 \tag{6}$$

计算 u 的算法可以如下进行：从最里边的括号开始，我们用 u_k，u_{k-1}，…，u_0 表示括号中的量. 因此

$$\begin{aligned} u_k &\leftarrow c_k \\ u_{k-1} &\leftarrow u_k d_{k-1} + c_{k-1} \\ u_{k-2} &\leftarrow u_{k-1} d_{k-2} + c_{k-2} \\ &\vdots \\ u_0 &\leftarrow u_1 d_0 + c_0 \end{aligned}$$

因为只需要 u_0，我们可以用算法形式写成

```
u ← c_k
for i = k-1 to 0 step -1 do
    u ← u d_i + c_i
end do
```

310

返回到(3)式或(4)式给出的多项式，对于 t 的给定值，用下面的算法得到 $u = p_k(t)$：

```
u ← c_k
for i = k-1 to 0 step -1 do
  u ← (t - x_i)u + c_i
end do
```

现在我们可以写出一个计算(4)式中系数 c_i 的算法. 可以看出(2)式中的系数 c 恰好就是后面记作的 c_k. 于是，计算 c_k 的公式是

$$c_k = \frac{y_k - p_{k-1}(x_k)}{(x_k - x_0)(x_k - x_1)\cdots(x_k - x_{k-1})}$$

从 x_0，x_1，…，x_n 和 y_0，y_1，…，y_n 的表值计算 c_0，c_1，…，c_n 的算法如下：

```
c0 ← y0
for k=1 to n do
  d ← xk − xk−1
  u ← ck−1
  for i=k−2 to 0 step −1 do
      u ← u(xk − xi) + ci
      d ← d(xk − xi)
  end do
  ck ← (yk − u)/d
end do
```

内循环计算 $p_{k-1}(x_k)$ 和 $(x_k-x_0)(x_k-x_1)\cdots(x_k-x_{k-1})$，它是由前面讨论过的算法直接修正而来的.

刚才给出的算法仅仅是作为教学用的. 它说明把一个构造性的存在性证明转为适合于计算机程序的算法过程是重要的. 然而，还有一个更为有效的方法可以得到与上面相同的结果. 这个方法是用均差来计算(4)式中的系数 c_0，c_1，…，c_k. 我们将在 6.2 节中介绍它.

对前面的算法编程并给出简单的检验如下：我们定义多项式

$$p_3(x) = 4x^3 + 35x^2 - 84x - 954 \tag{7}$$

下面给出这个函数的 4 个值：

(8)

x	5	−7	−6	0
y	1	−23	−54	−954

把给定的这个数值集合输入计算机程序后，可以正确地计算出系数 $c_0=1$，$c_1=2$，$c_2=3$，$c_3=4$. 这些系数是(7)式中牛顿型插值多项式的系数；即

$$p_3(x) = 1 + 2(x-5) + 3(x-5)(x+7) + 4(x-5)(x+7)(x+6)$$

311

6.1.2 拉格朗日型插值多项式

我们现在介绍另一种与一组给定的数据点 (x_i, y_i)，$0 \leqslant i \leqslant n$，联系在一起的插值多项式 p. 重要的是要理解这样的事实：存在并且仅存在一个与这组数据（当然，假设 $n+1$ 个横坐标 x_i 是不同的）联系在一起的次数不超过 n 次的插值多项式. 然而，这个多项式确实可能存在不同的表示形式，并且也可以用不同的算法来求得此多项式.

这种方法是把多项式 p 表成下列形式：

$$p(x) = y_0\ell_0(x) + y_1\ell_1(x) + \cdots + y_n\ell_n(x) = \sum_{k=0}^{n} y_k\ell_k(x) \tag{9}$$

其中 ℓ_0，ℓ_1，…，ℓ_n 是一些只依赖于结点 x_0，x_1，…，x_n 而与纵坐标 y_0，y_1，…，y_n 无关的多项式. 因为所有的纵坐标除了第 i 个分量是 1 之外，其他分量都是 0，所以我们有

$$\delta_{ij} = p_n(x_j) = \sum_{k=0}^{n} y_k\ell_k(x_j) = \sum_{k=0}^{n} \delta_{ki}\ell_k(x_j) = \ell_i(x_j)$$

(回忆**克罗内克 δ** 函数的定义是：当 $k=i$ 时 $\delta_{ki}=1$，当 $k\neq i$ 时 $\delta_{ki}=0$.) 我们不难得到一组具有这种性质的多项式.

我们先考察 ℓ_0，它是一个 n 次多项式，在点 x_1，x_2，…，x_n 处的值是 0，而在点 x_0 处的值是 1. 显然，ℓ_0 具有形式

$$\ell_0(x)=c(x-x_1)(x-x_2)\cdots(x-x_n)=c\prod_{j=1}^{n}(x-x_j)$$

令 $x=x_0$，由于

$$1=c\prod_{j=1}^{n}(x_0-x_j)$$

可得到 c 的值

$$c=\prod_{j=1}^{n}(x_0-x_j)^{-1}$$

因此，我们有

$$\ell_0(x)=\prod_{j=1}^{n}\frac{x-x_j}{x_0-x_j}$$

同理，可以得到每一个 ℓ_i，它的一般形式是

$$\ell_i(x)=\prod_{\substack{j=0\\ j\neq i}}^{n}\frac{x-x_j}{x_i-x_j}\qquad(0\leqslant i\leqslant n)\tag{10}$$

对于结点 x_0，x_1，…，x_n 的集合，这些多项式被称为**基函数**. 而用它们，(9)式给出了**拉格朗日型插值多项式**.

312

例 1 数表(8)中数据的基函数以及拉格朗日型插值多项式是什么？

解 因为结点是 5，−7，−6，0，所以基函数是

$$\ell_0(x)=\frac{(x+7)(x+6)x}{(5+7)(5+6)5}=\frac{1}{660}\times(x+6)(x+7)$$

$$\ell_1(x)=\frac{(x-5)(x+6)x}{(-7-5)(-7+6)(-7)}=\frac{-1}{84}\times(x-5)(x+6)$$

$$\ell_2(x)=\frac{(x-5)(x+7)x}{(-6-5)(-6+7)(-6)}=\frac{-1}{66}\times(x-5)(x+7)$$

$$\ell_3(x)=\frac{(x-5)(x+7)(x+6)}{(0-5)(0+7)(0+6)}=\frac{-1}{210}(x-5)(x+6)(x+7)$$

拉格朗日型插值多项式是

$$p_3(x)=\ell_0(x)-23\ell_1(x)-54\ell_2(x)-954\ell_3(x)$$

这个表达式看上去与(7)式中的 p_3 不同. 尽管它们的表示形式不同，但作为一个函数来说它们是恒等的. ■

例 2 求出下列两点数表中的插值公式.

x	x_0	x_1
y	y_0	y_1

解 拉格朗日插值公式为

$$p(x) = y_0\left(\frac{x-x_1}{x_0-x_1}\right)+y_1\left(\frac{x-x_0}{x_1-x_0}\right)$$ ■

还有其他一些插值多项式的算法，这些算法具有各自的优缺点．对于给定的 $n+1$ 个（不同的）数据点，因为存在且仅存在一个次数不超过 n 次的插值多项式，所以这些不同的算法产生具有不同形式的同一个多项式．例如，我们可以把多项式表示为 x 的方幂形式

$$p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n \tag{11}$$

为了确定 a_0，a_1，…，a_n，对于 $0\leqslant i\leqslant n$，由插值条件 $p(x_i)=y_i$ 产生了 $n+1$ 个线性方程的方程组，其形式是

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix}\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}=\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{12}$$

313

(12)式中的系数矩阵称为**范德蒙德矩阵**．由于对任意选取的 y_0，y_1，…，y_n 方程组有唯一解，所以该矩阵是非奇异的(定理1)．因此，对于不同结点 x_0，x_1，…，x_n 的范德蒙德矩阵的行列式非零，它的计算公式在习题 6.1.34 中给出．但是，*范德蒙德矩阵常常是病态的*，因此通过解(12)式所求出的 a_i 也可能不精确．（参见 Gautschi [1984]．）此外，得到(11)式中多项式的计算量也是非常大的．因此，我们并不建议采用这个方法．

如果令 $x=x_i$，其中 $0\leqslant i\leqslant n$，对于(9)式中的拉格朗日多项式，我们得到下列 $n+1$ 个线性方程：

$$\begin{bmatrix} \ell_0(x_0) & \ell_1(x_0) & \cdots & \ell_n(x_0) \\ \ell_0(x_1) & \ell_1(x_1) & \cdots & \ell_n(x_1) \\ \vdots & \vdots & \ddots & \vdots \\ \ell_0(x_n) & \ell_1(x_n) & \cdots & \ell_n(x_n) \end{bmatrix}\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}=\begin{bmatrix} p(x_0) \\ p(x_1) \\ \vdots \\ p(x_n) \end{bmatrix}$$

它的系数矩阵约化为单位矩阵 I，如前所述，它的解是 $y_i=p(x_i)$，其中 $0\leqslant i\leqslant n$．因为对 $n+1$ 个不同的点 x_0，x_1，…，x_n 只存在一个插值多项式，所以我们得到任意一个次数至多为 n 次的多项式所满足的恒等式

$$p(x) = \sum_{i=0}^{n} p(x_i)\ell_i(x)$$

就数值计算来说，最好采用牛顿型插值多项式．它可以结合均差算法（在 6.2 节中讨论）去求得所需系数．另一方面，由于拉格朗日插值公式中的系数是给定的 y_i，我们可以立即得到拉格朗日型插值多项式．在后面第 7 章构造求积公式时，这个事实将会很有用．

正如某些实验数据所反映的那样，如果一组固定的结点 x_i 对应多组不同的数值 y_i，因为对每一种情况基函数都保持不变，所以此时拉格朗日型插值优于牛顿型插值．

如果需要对插值问题添加新的数据点，那么牛顿型插值的另一个优点是已经算出的系数将保持不变．因此，在(3)式中，c_0 只与点(x_0, y_0)相关，c_1 只与点(x_0, y_0)和点(x_1, y_1)相关，

依此类推，牛顿型插值很容易适应于那些需要添加数据点的插值问题.

对牛顿型和(11)式型的计算可应用有效的嵌套乘法算法，这一点显然优于拉格朗日型插值. 但是，也有一些计算拉格朗日型插值的有效算法，参见 Werner [1984] 以及该文中的参考文献，也可参见习题 6.1.31～6.1.32.

6.1.3 多项式插值的误差

314

现在介绍一些有关函数和它的插值多项式之间差异的定理.

定理 2(*多项式插值误差定理*) *设 f 是 $C^{n+1}[a, b]$ 中的一个函数，多项式 p 是函数 f 在区间 $[a, b]$ 的 $n+1$ 个不同点 $x_0, x_1, \cdots, x_n$ 上的次数不超过 n 次的插值多项式. 对 $[a, b]$ 中的每个 x，都有 (a, b) 中的一点 ξ_x 与之对应，使得*

$$f(x) - p(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi_x) \prod_{i=0}^{n} (x - x_i) \tag{13}$$

证明 当 x 是某个插值结点 x_i 时，因为(13)式两端都为零，所以定理结论显然成立. 现在设 x 是异于结点的任意一点. 令

$$w(t) \equiv \prod_{i=0}^{n} (t - x_i) \qquad \phi \equiv f - p - \lambda w$$

其中 λ 是使得 $\phi(x)=0$ 的一个实数.（记住，x 是固定的.）因此，

$$\lambda = \frac{f(x) - p(x)}{w(x)}$$

现在 $\phi \in C^{n+1}[a, b]$，并且在 $n+2$ 个点 $x, x_0, x_1, \cdots, x_n$ 处变成零. 根据罗尔定理，ϕ' 在 (a, b) 中至少有 $n+1$ 个不同的零点. 类似地，ϕ'' 在 (a, b) 中至少有 n 个不同的零点. 如果重复上面的讨论，最终我们断定 $\phi^{(n+1)}$ 在 (a, b) 中至少有一个零点，例如 ξ_x. 因为

$$\phi^{(n+1)} = f^{(n+1)} - p^{(n+1)} - \lambda w^{(n+1)} = f^{(n+1)} - (n+1)!\lambda$$

因此，我们有

$$0 = \phi^{(n+1)}(\xi_x) = f^{(n+1)}(\xi_x) - (n+1)!\lambda = f^{(n+1)}(\xi_x) - (n+1)!\frac{f(x) - p(x)}{w(x)}$$

这是(13)式的变形. ■

例 3 设函数 $f(x)=\sin x$. 如果用函数 f 在区间[0, 1]中 10 个点上的 9 次插值多项式去逼近 f，试问在该区间上它们的误差有多大?

解 为了回答这个问题，我们要用到定理 2 中的(13)式. 显然，我们有 $|f^{(10)}(\xi_x)| \leqslant 1$ 和 $\prod_{i=0}^{9} |x - x_i| \leqslant 1$. 因此，对[0, 1]中的所有 x，

$$|\sin x - p(x)| \leqslant \frac{1}{10!} < 2.8 \times 10^{-7}$$ ■

6.1.4 切比雪夫多项式

通过适当地选取结点，定理 2 中有一项可以被优化. 这个问题的分析最初是由伟大的数学家切比雪夫(1821－1894)给出的. 优化的过程自然地引导出了一个称为**切比雪夫多项式**的一组多项式，下面给出它的定义及基本性质.

315

切比雪夫多项式(第一类)递归定义如下：

$$\begin{cases} T_0(x) = 1 \qquad T_1(x) = x \\ T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \qquad (n \geqslant 1) \end{cases}$$

不难算出接下来几个 T_n 的具体形式：

$$T_2(x) = 2x^2 - 1$$
$$T_3(x) = 4x^3 - 3x$$
$$T_4(x) = 8x^4 - 8x^2 + 1$$
$$T_5(x) = 16x^5 - 20x^3 + 5x$$
$$T_6(x) = 32x^6 - 48x^4 + 18x^2 - 1$$

图 6-1 给出了这最初几个切比雪夫多项式的图形.

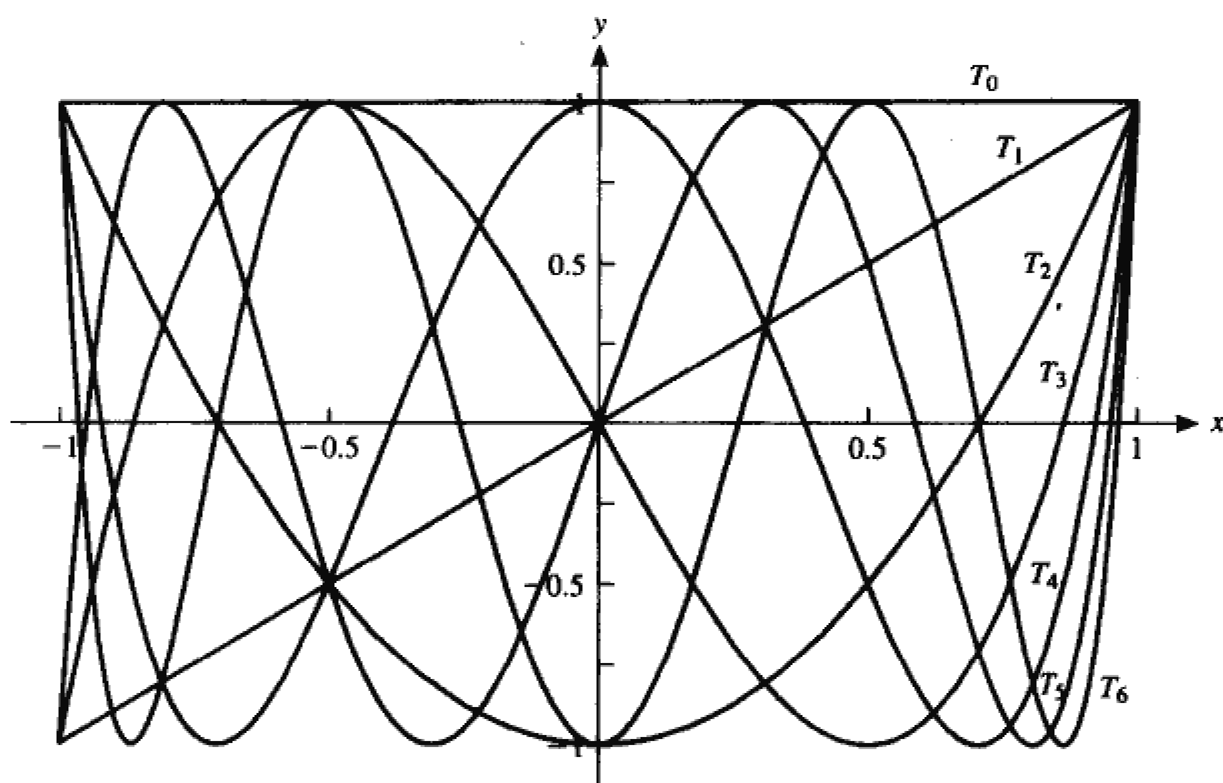


图 6-1 切比雪夫多项式 $T_n(x)$，$0 \leqslant n \leqslant 6$

切比雪夫(Чебышев)在研究蒸汽机车联动装置的运动时引出了这些多项式. 从那时起，它们就成为应用数学的重要组成部分. 切比雪夫多项式有许多绝妙的性质.(参见 Rivlin [1990].)

定理 3(切比雪夫多项式定理) 对区间[-1，1]中的 x，切比雪夫多项式有闭型表达式

$$T_n(x) = \cos(n \cos^{-1} x) \qquad (n \geqslant 0)$$

证明 回忆两角和的余弦公式是

$$\cos(A + B) = \cos A \cos B - \sin A \sin B$$

316

由此得到

$$\cos(n+1)\theta = \cos\theta \cos n\theta - \sin\theta \sin n\theta$$
$$\cos(n-1)\theta = \cos\theta \cos n\theta + \sin\theta \sin n\theta$$

两式相加后再移项，得到

$$\cos(n+1)\theta = 2\cos\theta\cos n\theta - \cos(n-1)\theta \tag{14}$$

现在令 $\theta=\cos^{-1}x$ 以及 $x=\cos\theta$. (14)式表明下列定义的函数

$$f_n(x) = \cos(n\cos^{-1}x)$$

满足

$$\begin{cases} f_0(x) = 1 \qquad f_1(x) = x \\ f_{n+1}(x) = 2xf_n(x) - f_{n-1}(x) \qquad (n \geqslant 1) \end{cases}$$

因此对所有的 n，$f_n=T_n$ 都成立. ■

根据定理 3 中的公式，我们可以得到切比雪夫多项式更多的性质，例如

$$|T_n(x)| \leqslant 1 \qquad (-1 \leqslant x \leqslant 1)$$

$$T_n\left(\cos\frac{j\pi}{n}\right) = (-1)^j \qquad (0 \leqslant j \leqslant n)$$

$$T_n\left(\cos\frac{2j-1}{2n}\pi\right) = 0 \qquad (1 \leqslant j \leqslant n)$$

首一多项式是指最高次项系数是 1 的一个多项式. 根据切比雪夫多项式的定义，对于 $n>0$，我们可以看出 $T_n(x)$中最高次项是 $2^{n-1}x^n$. 因此，当 $n>0$ 时，$2^{1-n}T_n$ 是一个首一多项式.

定理 4(首一多项式定理) *若 p 是一个 n 次首一多项式，则*

$$\|p\|_\infty = \max_{-1\leqslant x\leqslant 1} |p(x)| \geqslant 2^{1-n}$$

证明 用反证法证明. 假设

$$|p(x)| < 2^{1-n} \qquad (|x| \leqslant 1)$$

令 $q=2^{1-n}T_n$ 并且 $x_i=\cos(i\pi/n)$. 由前面的讨论知，q 是一个 n 次的首一多项式. 因而

$$(-1)^i p(x_i) \leqslant |p(x_i)| < 2^{1-n} = (-1)^i q(x_i)$$

并且，

317

$$(-1)^i[q(x_i) - p(x_i)] > 0 \qquad (0 \leqslant i \leqslant n)$$

这说明在区间[-1, 1]上，多项式 $q-p$ 的符号在正负之间来回变动了 $n+1$ 次. 所以它在(-1, 1)中至少有 n 个根. 但是这显然是不可能的，因为 $q-p$ 的次数至多是 $n-1$ (记住 p 和 q 都是首一多项式，因而 n 次项 x^n 在 $q-p$ 中绝不会出现). ■

6.1.5 选取结点

在定理 2 中，假设插值结点都在区间[-1, 1]内. 如果 x 在同一区间内，那么 ξ_x 也在该区间内. 因此，我们可得到

$$\max_{|x|\leqslant 1} |f(x) - p(x)| \leqslant \frac{1}{(n+1)!} \max_{|x|\leqslant 1} |f^{(n+1)}(x)| \max_{|x|\leqslant 1} \left|\prod_{i=0}^{n}(x-x_i)\right|$$

根据定理 4，我们有(对任何结点集)

$$\max_{|x|\leqslant 1} \left|\prod_{i=0}^{n}(x-x_i)\right| \geqslant 2^{-n}$$

若 $\prod_{i=0}^{n}(x-x_i)$ 是首一多项式 $2^{-n}T_{n+1}$，则上述不等式可以达到极小值. 因而结点就是 T_{n+1} 的根，它们是

$$x_i = \cos\left(\frac{2i+1}{2n+2}\pi\right) \qquad (0 \leqslant i \leqslant n)$$

综上所述，得到下列结果

定理 5(切比雪夫结点插值误差定理) 若结点 x_i 是切比雪夫多项式 T_{n+1} 的根，则对 $|x| \leqslant 1$，定理 2 中的误差公式变为

$$|f(x) - p(x)| \leqslant \frac{1}{2^n(n+1)!} \max_{|t| \leqslant 1} |f^{(n+1)}(t)|$$

6.1.6 插值多项式的收敛性

设 f 是定义在区间$[a, b]$上的一个连续函数，如果对 f 构造次数越来越高（具有等分结点）的插值多项式 p_n，那么自然会期望这些多项式在$[a, b]$上一致收敛于 f，也就是当 $n \to \infty$ 时，我们期望

$$\| f - p_n \|_\infty = \max_{a \leqslant x \leqslant b} |f(x) - p_n(x)|$$

收敛于零．我们已经看到过一些如 $f(x) = \sin x$ 那样的函数例子，其上述结论是正确的．但是必须意识到，函数 $\sin x$ 远不是一个典型的连续函数．在实数域上它是属于 C^∞ 类的函数，而作为复变函数，它又是一个**整函数**．也就是说它在复平面(的有限区域)上根本没有奇点． 318

出人意料的情形是对于大部分连续函数而言，$\| f - p_n \|_\infty$ 并不收敛于零．这种情形的第一个例子由 Meray 于 1884 年给出．他选择复平面上的 n 个 n **次单位根**作为结点，它们是单位圆周 $|z| = 1$ 上的等分结点．当 $n=6$ 时如图 6-2 所示．

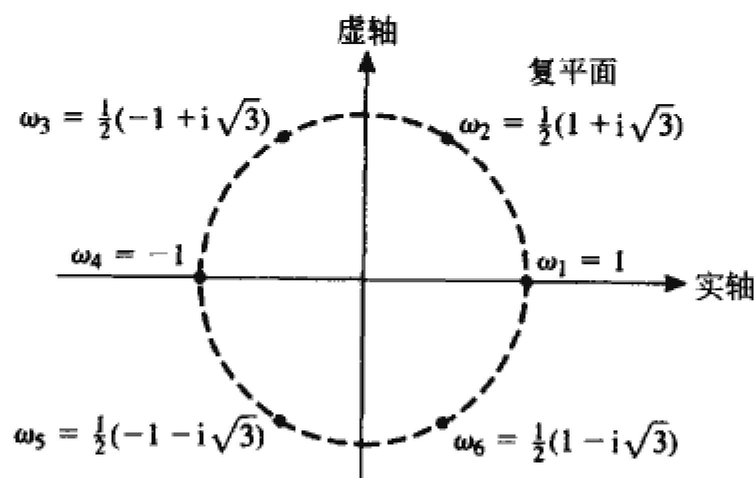


图 6-2 6 个 6 次单位根

我们把这些结点记为 $\omega_1, \omega_2, \cdots, \omega_n$．现在考虑函数 $f(z) = 1/z$ 在结点 $\omega_1, \omega_2, \cdots, \omega_n$ 上的多项式插值问题．这个问题的解是唯一的 $n-1$ 次的多项式 p_{n-1}．毫无疑问，它是 $p_{n-1}(z) = z^{n-1}$，这是因为

$$p_{n-1}(\omega_j) = \omega_j^{n-1} = \frac{1}{\omega_j}\omega_j^n = \frac{1}{\omega_j} = f(\omega_j) \qquad (1 \leqslant j \leqslant n)$$

我们来度量在单位圆上 f 和 p_{n-1} 之间的差异．我们有

$$\| f - p_{n-1} \|_\infty = \max_{|z|=1} |f(z) - p_{n-1}(z)| = \max_{|z|=1} |z^{-1} - z^{n-1}|$$

$$= \max_{|z|=1} \frac{1}{|z|} |1 - z^n| = 2$$

(当 z 在单位圆上 $|z|=1$ 移动时，z^n 也在其上移动．并且当 $z^n=-1$ 时，由上述计算我们得到 2.）因此，当 $n \to \infty$ 时，此例中的 p_n 和 f 总是相距两个单位.

另一个例子是在实数域内由龙格于 1901 年给出的．函数 $f(x)=(x^2+1)^{-1}$ 定义在区间 $[-5, 5]$ 上．如果 p_n 是由这个函数利用区间 $[-5, 5]$ 内的等距结点构造的插值多项式，我们发现序列 $\|f-p_n\|_\infty$ 是无界的．从下一节的计算机实验中可以看到这种情形的轮廓．即使对不太大的 n（例如 $n=15$），我们也可以看到多项式 p_n 的剧烈振荡．这个奇妙的现象已经成为教科书和论文中的标准例题，被称为**龙格函数**或者**龙格例题**，参见 Epperson[1987].

关于龙格函数在等距结点上插值产生无界多项式的证明，可参见 Steffensen [1950]关于插值的一本专著，该书于 1927 年第一次出版．进一步的背景情况需要用到复函数来处理．为此，我们推荐 Davis [1982] 有关逼近和插值方面的专著．以实函数的观点来看，龙格函数似乎具有良好的性质．但是在复平面内，它在虚轴附近有奇点．这便是麻烦的根源所在，对它的分析相当困难并且超出了本书的范围．一个肯定的结论是下面的定理 7，它表明在某种条件下插值多项式一致收敛于 f．然而，一个否定的结论是下面的定理 6，它表明对应于 f 的插值多项式是无界的(因此不能一致收敛于 f)．这两个定理的不同之处就在于它们的量词.

319

接着是其他一些数学家的工作．例如，法贝尔(Faber)在 1914 年发现了下面极具一般性的结果：

定理 6(法贝尔定理)　*对任意给定的结点组*

$$a \leqslant x_0^{(n)} < x_1^{(n)} < \cdots < x_n^{(n)} \leqslant b \qquad (n \geqslant 0) \tag{15}$$

在区间 $[a, b]$ 上存在一个连续函数 f，使得 f 在这组结点上的插值多项式不能一致收敛于 f.

由于我们有下面肯定的结论，所以这个结果比初看上去要难以理解些.

定理 7(插值法收敛性定理)　*若 f 是 $[a, b]$ 上的连续函数，则存在 (15) 式中那样的一组结点，使得 f 在这组结点上的插值多项式 p_n 满足 $\lim_{n\to\infty}\|f-p_n\|_\infty=0$.*

这个定理可以由下面两个著名的定理结合起来得到：**魏尔斯特拉斯逼近定理和切比雪夫交错定理**．我们现在只给出其中之一．（另一个将在 6.9 节中给出.）

定理 8(魏尔斯特拉斯逼近定理)　*若 f 是 $[a, b]$ 上的连续函数并且 $\varepsilon>0$，则在区间 $[a, b]$ 上存在一个多项式 p 满足 $|f(x)-p(x)| \leqslant \varepsilon$.*

证明　由于变量替换 $x=a+t(b-a)$ 可以用来在 $[a, b]$ 和 $[0, 1]$ 之间转换，因此只需在特定的区间 $[0, 1]$ 上证明这个定理即可．当然，t 的线性函数插值多项式是 t 的同次多项式．如果 $f \in C[0, 1]$，正如后面将要解释的那样，伯恩斯坦(Bernstein)多项式 $B_n f$ 序列一致收敛于 f. ■

现在，我们讨论 Serge Bernstein(伯恩斯坦)在 1912 年引进的**伯恩斯坦多项式**．它由下式给出：

320

$$(B_n f)(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) g_{nk}(x) \qquad 其中 \qquad g_{nk}(x) = \binom{n}{k} x^k (1-x)^{n-k}$$

在 g_{nk} 的定义中，记号 $\binom{n}{k}$ 是由下式定义的**二项式系数**

$$\binom{n}{k}=\begin{cases}\dfrac{n!}{k!(n-k)!} & 0\leqslant k\leqslant n\\ 0 & \text{其他}\end{cases}$$

为什么我们要关注伯恩斯坦多项式？我们主要是用它们来给出魏尔斯特拉斯定理的一个初等证明．然而，伯恩斯坦多项式也被用于计算机辅助设计．（最近，在这个应用领域里有一种用 B 样条替代这些多项式的趋势．）

把 B_n 看作为 $C[0,1]$中的元素映射到另一些元素的线性算子，这一点是很重要的．B_n 的线性性可由下列等式表示：

$$B_n(af+bg)=aB_nf+bB_ng \qquad (a,b\in\mathbb{R};f,g\in C[0,1])$$

这一点很容易验证．这些算子的另一个重要性质是它们是正的，也就是说如果 $f\geqslant 0$，那么 $B_nf\geqslant 0$．根据在区间$[0,1]$上 $g_{nk}\geqslant 0$ 的事实，易知这一点是成立的．（我们把所有函数都看作定义在区间$[0,1]$上．）

正如下面定理所指出的那样，算子的线性性和正性是非常有效的．

定理 9（Bohman-Korovkin 定理） *设 $L_n(n\geqslant 1)$是定义在 $C[a,b]$上的一个正线性算子序列并且其在相同的空间里取值．若对于三个函数 $f(x)=1,x,x^2$，$\|L_nf-f\|_\infty\to 0$ 都成立，则对所有 $f\in C[a,b]$此结论也成立．*

证明 在空间 $C[a,b]$中，我们可以取绝对值，用 $|f|$ 表示在 x 处的值是 $|f(x)|$ 的函数．我们还注意到，如果 L 是正线性算子，则有下列关系

$$f\geqslant g\Rightarrow f-g\geqslant 0\Rightarrow L(f-g)\geqslant 0\Rightarrow Lf-Lg\geqslant 0\Rightarrow Lf\geqslant Lg$$

因为 $|f|\geqslant f$ 和 $|f|\geqslant -f$，所以 $L(|f|)\geqslant Lf$ 和 $L(|f|)\geqslant -Lf$，进一步有 $L(|f|)\geqslant |Lf|$．对于 $k=0,1,2$，我们设 $h_k(x)=x^k$，再定义函数 α_n，β_n 和 γ_n 如下

$$\alpha_n=L_nh_0-h_0 \qquad \beta_n=L_nh_1-h_1 \qquad \gamma_n=L_nh_2-h_2$$

由定理假设可以断定

$$\|\alpha_n\|_\infty\to 0 \qquad \|\beta_n\|_\infty\to 0 \qquad \|\gamma_n\|_\infty\to 0$$

（在此处以及定理证明的剩余部分中，我们使用了无穷范数．）在定理证明的主要部分，设 f 是 $C[a,b]$中的任意元素并且 $\varepsilon>0$．我们要证明存在一个整数 m 使得

321

$$n\geqslant m\Rightarrow\|L_nf-f\|_\infty<3\varepsilon$$

由于 f 在一个紧区间上连续，从而它一致连续．因此，存在一个正数 δ，使得对于区间$[a,b]$中所有的 x 和 y，都有

$$|x-y|<\delta\Rightarrow|f(x)-f(y)|<\varepsilon$$

就 $c=2\|f\|_\infty/\delta^2$，我们有

$$|x-y|\geqslant\delta\Rightarrow|f(x)-f(y)|\leqslant 2\|f\|_\infty\leqslant 2\|f\|_\infty\frac{(x-y)^2}{\delta^2}=c(x-y)^2$$

从而，对于$[a,b]$中所有的 x 和 y，我们有

$$|f(x)-f(y)|\leqslant\varepsilon+c(x-y)^2$$

这个不等式可写为下列形式

$$|f-f(y)h_0|\leqslant\varepsilon h_0+c[h_2-2yh_1+y^2h_0]$$

根据 h_k 的定义以及简单地代换 x 可以得到上述结果．由定理证明开始时所作的注记，我们得到

$$|L_nf-f(y)L_nh_0|\leqslant \varepsilon L_nh_0+c[L_nh_2-2yL_nh_1+y^2L_nh_0]$$

这是一个函数之间的不等式，我们可以随意地将 y 代入其中：

$$\begin{aligned}
&|(L_nf)(y)-f(y)(L_nh_0)(y)|\\
&\leqslant \varepsilon(L_nh_0)(y)+c[(L_nh_2)(y)-2y(L_nh_1)(y)+y^2(L_nh_0)(y)]\\
&=\varepsilon[1+\alpha_n(y)]+c[y^2+\gamma_n(y)-2y(y+\beta_n(y))+y^2(1+\alpha_n(y))]\\
&=\varepsilon+\varepsilon\alpha_n+c\gamma_n(y)-2cy\beta_n(y)+cy^2\alpha_n(y)\\
&\leqslant \varepsilon+\varepsilon\|\alpha_n\|_\infty+c\|\gamma_n\|_\infty+2c\|h_1\|_\infty\|\beta_n\|_\infty+c\|h_2\|_\infty\|\alpha_n\|_\infty
\end{aligned}$$

选择 m，使得当 $n\geqslant m$ 时，上述不等式中最后的右端项小于 2ε．那么对于 $n\geqslant m$，我们有

$$\|L_nf-f\cdot L_nh_0\|_\infty\leqslant 2\varepsilon$$

最后，我们有

322

$$\begin{aligned}
\|L_nf-f\|_\infty&\leqslant\|L_nf-f\cdot L_nh_0\|_\infty+\|f\cdot L_nh_0-f\cdot h_0\|_\infty\\
&\leqslant 2\varepsilon+\|f\|_\infty\|\alpha_n\|_\infty
\end{aligned}$$

如果需要的话，可以增加 m 使得当 $n\geqslant m$ 时，有 $\|f\|_\infty\|\alpha_n\|_\infty<\varepsilon$．那么上述不等式中的最后项不超过 3ε． ■

在魏尔斯特拉斯定理的证明中，遗留的一个细节是：对于 $k=0，1，2$，证明 $B_kh_k\to h_k$．其中 $h_k(x)=x^k$．对于 h_0，我们用二项式定理给出

$$(B_nh_0)(x)=\sum_{k=0}^n g_{nk}(x)=\sum_{k=0}^n\binom{n}{k}x^k(1-x)^{n-k}=(x+(1-x))^n=1$$

对于 h_1，根据习题 6.1.35，我们有

$$\begin{aligned}
(B_nh_1)(x)&=\sum_{k=0}^n\left[\frac{k}{n}\right]\binom{n}{k}x^k(1-x)^{n-k}=\sum_{k=1}^n\binom{n-1}{k-1}x^k(1-x)^{n-k}\\
&=x\sum_{k=0}^{n-1}\binom{n-1}{k}x^kx^{n-1-k}=x
\end{aligned}$$

最后，对于 $k=2$，两次应用习题 6.1.35，我们有

$$\begin{aligned}
(B_nh_2)(x)&=\sum_{k=0}^n\left[\frac{k}{n}\right]^2\binom{n}{k}x^k(1-x)^{n-k}=\sum_{k=1}^n\left[\frac{k}{n}\right]\binom{n-1}{k-1}x^k(1-x)^{n-k}\\
&=\sum_{k=1}^n\left[\frac{n-1}{n}\frac{k-1}{n-1}+\frac{1}{n}\right]\binom{n-1}{k-1}x^k(1-x)^{n-k}\\
&=\frac{n-1}{n}x^2\sum_{k=2}^n\binom{n-2}{k-2}x^{k-2}(1-x)^{n-k}+\frac{x}{n}\\
&=\frac{n-1}{n}x^2+\frac{x}{n}\to x^2
\end{aligned}$$

习题 6.1

1．求出下列数据集上次数最低的插值多项式：

a.

x	3	7
y	5	-1

b.

x	7	1	2
y	146	2	1

c.

x	3	7	1	2
y	10	146	2	1

d.

x	3	7	1	2
y	12	146	2	1

e.

x	1.5	2.7	3.1	−2.1	−6.6	11.0
y	0.0	0.0	0.0	1.0	0.0	0.0

2. 对于给定的函数 $f(x)$，在 $n+1$ 个指定的结点上次数$\leqslant n$ 的插值多项式 p 是唯一确定的. 因此，存在一个映射 $f\mapsto p$. 用 L 表示这个映射并且证明 323

$$Lf=\sum_{i=0}^{n}f(x_i)\ell_i$$

再证明 L 是线性的，即 $L(af+bg)=aLf+bLg$

3.（续）参考上题并且由如下公式定义另一个映射 G：

$$Gf=\sum_{i=0}^{n}f(x_i)\ell_i^2$$

证明 Gf 是一个次数至多是 $2n$ 次的多项式，Gf 是 f 在给定结点上的插值多项式. 当 f 非负时 Gf 也是非负的.

4.（续）证明上面习题 6.1.2 中的映射 L 具有性质：对每个次数最多是 n 次的多项式 q，都有 $Lq=q$.

5.（续）证明对所有 x，都有 $\sum_{i=0}^{n}\ell_i(x)=1$.

6.（续）证明：如果 p 是函数 f 在结点 x_0，x_1，…，x_n（不同的点）上次数$\leqslant n$ 的插值多项式，那么

$$f(x)-p(x)=\sum_{i=0}^{n}[f(x)-f(x_i)]\ell_i(x)$$

7. 证明：牛顿型插值多项式中计算系数 c_i 的算法包含 n^2 次长运算（乘法和除法）.

8. 对于给定的 $p(0)$，$p(1)$和 $p'(\xi)$，其中 ξ 是任意预先给定的点. 讨论求一个次数至多为 2 次的多项式问题.

9. 证明：如果 g 是函数 f 在点 x_0，x_1，…，x_{n-1} 上的插值并且 h 是 f 在点 x_1，x_2，…，x_n 上的插值，则函数

$$g(x)+\frac{x_0-x}{x_n-x_0}[g(x)-h(x)]$$

是 f 在点 x_0，x_1，…，x_n 上的插值. 注意：g 和 h 不一定是多项式.

10. 证明：在(9)式的多项式 p 中 x^n 的系数是

$$\sum_{i=0}^{n}y_i\prod_{\substack{j=0\\j\neq i}}^{n}(x_i-x_j)^{-1}$$

11. 证明：对任何次数$\leqslant n-1$ 的多项式 q，

$$\sum_{i=0}^{n}q(x_i)\prod_{\substack{j=0\\j\neq i}}^{n}(x_i-x_j)^{-1}=0$$

12. 确定下列算法

```
x←a_n b_n
for i=1 to n do
```

```
  x←(x+a_{n-i})b_i
end do
```

是否可以计算

[324]

$$x = \sum_{i=0}^{n} a_i \prod_{j=0}^{i} b_j$$

13. 证明：如果我们在区间[-1，1]内任取 23 个结点并且 p 是函数 $f(x)=\cosh x$ 在结点上的一个 22 次插值多项式，则在[-1，1]上，相对误差 $|p(x)-f(x)|/|f(x)|$ 不大于 5×10^{-16}.

14. 设 p 是函数 $f(x)=\sinh x$ 在区间[-1，1]内任意 n 个结点上的一个次数 $\leqslant n-1$ 的插值多项式，仅有的限制条件是其中一个结点为 0. 证明在[-1，1]上误差满足下列不等式

$$|p(x)-f(x)| \leqslant \frac{2^n}{n!}|f(x)|$$

15. 试问下列算法中 v 的最后的值是多少？

```
v←c_{i-1}
for j=i to n do
  v←vx+c_j
end do
```

这个算法中共包含多少次加法和减法？

16. 写出一个有效的算法计算

$$u = \sum_{i=1}^{n} \prod_{j=1}^{i} d_j$$

17. 假设 p 是函数 f 在 $n+1$ 个结点上的一个次数大于 n 次的插值多项式. 关于 $f(x)-p(x)$ 你能发现什么？

18. 证明或者否定：如果 n 是 m 的因子，则 T_n 的每个零点也是 T_m 的零点.

19. 试求满足下面表值的一个多项式：

x	1	2	0	3
y	3	2	−4	5

20. 证明：对于 $x\geqslant 1$，$T_n(x)=\cosh(n\cosh^{-1}x)$. 提示：模仿关于切比雪夫多项式定理 3 的证明.

21. 写出关于下面表值的拉格朗日和牛顿插值多项式：

x	2	0	3
$f(x)$	11	7	28

22. 求下列数据的拉格朗日型和牛顿型插值多项式：

x	−2	0	1
$f(x)$	0	1	−1

[325] 把两个多项式都写成 $a+bx+cx^2$ 的形式并证明作为函数它们是恒等的.

23. 考察数据

x	$-\sqrt{\frac{3}{5}}$	0	$\sqrt{\frac{3}{5}}$
$f(x)$	$f\left(-\sqrt{\frac{3}{5}}\right)$	$f(0)$	$f\left(\sqrt{\frac{3}{5}}\right)$

关于这组数据的拉格朗日插值多项式和牛顿插值多项式是怎样的？

24. T_n 中的首项系数是 2^{n-1}，x^{n-2} 系数的公式是怎样的？x^{n-1} 的系数公式呢？

25. 求出下列数表的插值多项式：

x	1	3	2	6
y	-2	-22	-1	-37

26. 方程 $x-9^{-x}=0$ 在[0，1]中有一个解．给出该方程左端的函数在点 $x_0=0$，$x_1=0.5$，$x_2=1$ 上的插值多项式．令插值多项式等于零并且解这个方程，求出此方程的一个近似解．

27. 如果 p 是函数 $f(x)=e^{x-1}$ 在[-1，1]内的 13 个结点上的 12 次插值多项式，那么 $|f(x)-p(x)|$ 在[-1，1]上一个好的上界是什么？

28. 对于 $0\leqslant i\leqslant k$，设 p_k 是次数 $\leqslant k$ 使得 $p_k(x_i)=y_i$ 的多项式．证明：$p_k=p_{k-1}$ 当且仅当 $p_{k-1}(x_k)=y_k$．

29. 如果 f^{-1} 是所求根的某个邻域内的 n 次多项式，设计一个解方程 $f(x)=0$ 并且在 $n+1$ 步给出其精确根的算法．这里 f^{-1} 是反函数：$f^{-1}(f(x))=x$．

30. 证明：若 g 是函数 f 在结点 x_0，x_1，…，x_{n-1} 上的插值函数(不必是一个多项式)并且函数 h 满足 $h(x_i)=\delta_{in}(0\leqslant i\leqslant n)$，则对某个 c 函数 $g+cf$ 是 f 在 x_0，x_1，…，x_n 上的插值．

31. 参考拉格朗日插值过程并且定义

$$w_i=\prod_{\substack{j=0\\j\neq i}}^{n}(x_i-x_j)^{-1}$$

证明：若 x 不是结点，则插值多项式可由下列公式计算：

$$p(x)=\left[\sum_{i=0}^{n}y_iw_i(x-x_i)^{-1}\right]\Big/\left[\sum_{i=0}^{n}w_i(x-x_i)^{-1}\right]$$

这个公式称为拉格朗日插值过程的**重心形式**．

32. (续)证明在上面习题中 p 的表达式在下述意义下是稳定的：若 w_i 不能精确地计算出来，但仍然具有插值性质，即 $\lim_{x\to x_k}p(x)=y_k(0\leqslant k\leqslant n)$．

33. 设 E 是定义在域 D 上的函数所构成的 $n+1$ 维向量空间．设 x_0，x_1，…，x_n 是 D 中不同的点．证明插值问题

$$f(x_i)=y_i\qquad(0\leqslant i\leqslant n)\qquad f\in E$$

对任意选取的纵坐标 y_i 有唯一的解当且仅当 E 中没有异于 0 的函数在所有点 x_0，x_1，…，x_n 上值都是零． [326]

34. 证明

$$\det\begin{vmatrix}1 & x_0 & x_0^2 & \cdots & x_0^n\\ 1 & x_1 & x_1^2 & \cdots & x_1^n\\ 1 & x_2 & x_2^2 & \cdots & x_2^n\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & x_n & x_n^2 & \cdots & x_n^n\end{vmatrix}=\prod_{0\leqslant j<k\leqslant n}(x_k-x_j)$$

$$=(x_n-x_0)(x_n-x_1)(x_n-x_2)\cdots(x_n-x_{n-1})\cdots$$
$$(x_3-x_0)(x_3-x_1)(x_3-x_2)(x_2-x_0)$$
$$(x_2-x_1)(x_1-x_0)$$

35. 证明

$$\left[\frac{k}{n}\right]\binom{n}{k}=\binom{n-1}{k-1}$$

36. 函数 $1/(1+x^2)$ 和函数 e^{-x^2} 具有相似的图像．那么它们对等距结点的插值过程也是相似的吗？

37. 美国的第一张邮票发行于 1885 年，邮寄一封信的费用是 2 美分．1917 年，费用增加到 3 美分．但是，在

1919 年又降到 2 美分. 1932 年，其费用再次上升到 3 美分并且维持了 26 年. 然后，发生了一连串的增加过程：1958 年 4 美分，1963 年 5 美分，1968 年 6 美分，1971 年 8 美分，1974 年 10 美分，1978 年 15 美分，1981 年 3 月 18 美分及 10 月 20 美分，1985 年 22 美分，1988 年 25 美分，1991 年 29 美分，1995 年 32 美分，1999 年 33 美分，2001 年 34 美分. 确定对于这些数据的牛顿插值公式. 试问在此基础上，什么时候邮寄一封信要花费 1 美元？什么时候要花费 10 美元？

6.2 均差

在上节中，我们讨论了函数用多项式插值的问题，现在继续讨论这个问题. 设 f 是一个函数，其在点(结点)集 x_0，x_1，…，x_n 上的函数值是已知的或者可以计算出来. 本节中假设这些结点互不相同，但不要求它们按大小关系在实轴上有序地排列. 我们知道在这 $n+1$ 个结点上存在唯一一个次数至多是 n 次的 f 的插值多项式 p.

$$p(x_i) = f(x_i) \qquad (0 \leqslant i \leqslant n) \tag{1}$$

当然，多项式 p 可以写成基本多项式 1，x，x^2，…，x^n 的线性组合. 正如上节所讨论的那样，这些基本多项式不太令人满意，我们更喜欢使用合乎牛顿插值多项式的基本性质：

327

$$\begin{aligned}
q_0(x) &= 1 \\
q_1(x) &= (x - x_0) \\
q_2(x) &= (x - x_0)(x - x_1) \\
q_3(x) &= (x - x_0)(x - x_1)(x - x_2) \\
&\vdots \\
q_n(x) &= (x - x_0)(x - x_1)(x - x_2)\cdots(x - x_{n-1})
\end{aligned}$$

这些等式引导出牛顿插值多项式

$$p(x) = \sum_{j=0}^{n} c_j q_j(x)$$

插值条件给出了一个计算未知系数 c_j 的线性方程组

$$\sum_{j=0}^{n} c_j q_j(x_i) = f(x_i) \qquad (0 \leqslant i \leqslant n) \tag{2}$$

这个方程组中系数矩阵 A 是$(n+1)\times(n+1)$的，它的元素是

$$a_{ij} = q_j(x_i) \qquad (0 \leqslant i,j \leqslant n) \tag{3}$$

矩阵 A 是一个下三角阵，因为

$$q_j(x) = \prod_{k=0}^{j-1}(x - x_k)$$

$$q_j(x_i) = \prod_{k=0}^{j-1}(x_i - x_k) = 0 \qquad 若\ i \leqslant j-1 \tag{4}$$

例如，考虑 3 个结点的情形，并且

$$\begin{aligned}
p_2(x) &= c_0 q_0(x) + c_1 q_1(x) + c_2 q_2(x) \\
&= c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1)
\end{aligned}$$

设 $x=x_0$，$x=x_1$，$x=x_2$，我们有一个下三角方程组

$$\begin{bmatrix}1 & 0 & 0\\1 & (x_1-x_0) & 0\\1 & (x_2-x_0) & (x_2-x_0)(x_2-x_1)\end{bmatrix}\begin{bmatrix}c_0\\c_1\\c_2\end{bmatrix}=\begin{bmatrix}f(x_0)\\f(x_1)\\f(x_2)\end{bmatrix} \tag{5}$$

为求解(2)式中的 c_0，c_1，…，c_n，我们按照给定的下标顺序，从上往下计算系数 c_j. 在这个过程中，我们看到 c_0 只依赖于 $f(x_0)$，c_1 只依赖于 $f(x_0)$和 $f(x_1)$，依此类推. 因此，c_n 只依赖于 f 在点 x_0，x_1，…，x_n 上的值. 记号

328

$$c_n = f[x_0, x_1, \cdots, x_n] \tag{6}$$

在很多年以前就被用来表示这种依赖关系了. 因此，当 $\sum_{k=0}^{n} c_k q_k$ 是 f 在点 x_0，x_1，…，x_n 上的插值时，我们规定符号 $c_n=f[x_0,x_1,\cdots,x_n]$是 q_n 的系数. 因为

$$q_n(x) = (x-x_0)(x-x_1)\cdots(x-x_{n-1}) = x^n + \text{低次幂项}$$

所以我们也可以说 $f[x_0,x_1,\cdots,x_n]$是 f 在点 x_0，x_1，…，x_n 上次数至多为 n 次的插值多项式中 x^n 的系数. 在前面所有的描述中，n 可以取任意值. 表示式 $f[x_0,x_1,\cdots,x_n]$称为 f 的**均差**.

现在要给出前几个均差的显式公式. 首先，$f[x_0]$是 f 在点 x_0 上零次插值多项式中 x^0 的系数. 因此，我们有

$$f[x_0] = f(x_0) \tag{7}$$

$f[x_0,x_1]$是 f 在点 x_0，x_1 上次数至多为 1 次的插值多项式中 x 的系数. 因为这个多项式是

$$p(x) = f(x_0) + \frac{f(x_1)-f(x_0)}{x_1-x_0}(x-x_0) \tag{8}$$

所以我们看到 $q_1(x)$的系数是

$$f[x_0,x_1] = \frac{f(x_1)-f(x_0)}{x_1-x_0} \tag{9}$$

该表达式暗示出采用术语**均差**的原因. 具有下列形式的一个均差表可表示为

$$\begin{matrix} x_0 & f(x_0) & f[x_0,x_1] \\ x_1 & f(x_1) & \end{matrix}$$

由此很容易构造出插值多项式：

$$p(x) = f(x_0) + f[x_0,x_1](x-x_0)$$

由于 $c_0=f[x_0]$ 以及 $c_1=f[x_0,x_1]$与(6)式一致，因此(7)式和(9)式也可以通过求解方程组(5)中的 c_0 和 c_1 得到. 我们可以根据(1)式，把牛顿插值多项式写为如下形式：

$$p(x) = \sum_{k=0}^{n} c_k q_k(x) = \sum_{k=0}^{n} f[x_0,x_1,\cdots,x_k]\prod_{j=0}^{k-1}(x-x_j) \tag{10}$$

用下述理由可以说明(10)式的正确性：如果在(10)式的和中截取 $k=m$，得到 $\sum_{k=0}^{m} c_k q_k(x)$，我们知道这是 f 在点 $x_0,x_1,\cdots,x_m$ 上的次数至多是 m 次的插值多项式. 因此 $c_m=f[x_0,x_1,\cdots,x_m]$. 这一点对于 $0\leqslant m\leqslant n$ 中所有的 m 都是正确的.

6.2.1 高阶均差

下面的定理可以用来计算高阶均差.

329

定理 1(高阶均差定理)　均差满足等式

$$f[x_0,x_1,\cdots,x_n]=\frac{f[x_1,x_2,\cdots,x_n]-f[x_0,x_1,\cdots,x_{n-1}]}{x_n-x_0} \tag{11}$$

证明 首先，令 p_k 是 f 在结点 x_0，x_1，…，x_k 上次数至多是 k 次的插值多项式. 我们将用到 p_n 和 p_{n-1}. 用 q 表示 f 在点 x_1，x_2，…，x_n 上次数至多是 $n-1$ 次的插值多项式. 我们有

$$p_n(x)=q(x)+\frac{x-x_n}{x_n-x_0}[q(x)-p_{n-1}(x)] \tag{12}$$

下面证明这个等式成立. 首先，上式两端的多项式次数至多是 n 次的，其次，两端多项式在点 x_0，x_1，…，x_n 上的取值相同，因此，这两个多项式相等. 现在我们检查(12)式两端 x^n 的系数. 这两个系数必须相等，因此得到(11)式. ■

定理 1 给出了下列特殊的公式

$$f[x_0,x_1]=\frac{f[x_1]-f[x_0]}{x_1-x_0}$$

$$f[x_0,x_1,x_2]=\frac{f[x_1,x_2]-f[x_0,x_1]}{x_2-x_0}$$

$$\vdots$$

在这些公式中，x_0，x_1，x_2，…可以看作为独立变量，因此，我们有如下等式

$$f[x_i,x_{i+1},\cdots,x_{i+j}]=\frac{f[x_{i+1},x_{i+2},\cdots,x_{i+j}]-f[x_i,x_{i+1},\cdots,x_{i+j-1}]}{x_{i+j}-x_i} \tag{13}$$

其中 $f[x_i]$，$f[x_i, x_{i+1}]$，$f[x_i, x_{i+1}, x_{i+2}]$，$f[x_i, x_{i+1}, x_{i+2}, x_{i+3}]$，等等，分别是零阶，一阶，二阶，三阶等等均差. 当给定一个函数值$(x_i, f(x_i))$表时，我们可以构造出一个均差表. 这个表习惯上设计为下面的形式，在该表的每个相继列中给出零阶，一阶，二阶和三阶均差：

$$\begin{array}{ll|lll}
x_0 & f[x_0] & f[x_0, x_1] & f[x_0, x_1, x_2] & f[x_0, x_1, x_2, x_3] \\
x_1 & f[x_1] & f[x_1, x_2] & f[x_1, x_2, x_3] & \\
x_2 & f[x_2] & f[x_2, x_3] & & \\
x_3 & f[x_3] & & &
\end{array}$$

给定竖线左端的信息，并算出右端的量. 可用(11)式来实现这一点. (11)式的递归性质规定了均差表的三角形式，例如，给定的数据并不允许我们计算 $f[x_3, x_4]$，$f[x_2, x_3, x_4]$等. 330

比较一下(10)式和(11)式，我们看到牛顿插值多项式所需的系数位于均差表的第一行.

例 1 计算下列函数值的均差表：

x	3	1	5	6
$f(x)$	1	-3	2	4

(14)

解 把给定的数竖排为两列，用(11)式计算均差，得到

$$\begin{array}{cc|ccc}
3 & 1 & 2 & -\frac{3}{8} & \frac{7}{40} \\
1 & -3 & \frac{5}{4} & \frac{3}{20} & \\
5 & 2 & 2 & & \\
6 & 4 & & &
\end{array}$$

■

例 2 求出数表(14)中的函数值的牛顿插值多项式

解 $p(x)=1+2(x-3)-\frac{3}{8}(x-3)(x-1)+\frac{7}{40}(x-3)(x-1)(x-5)$ ■

6.2.2 均差的算法

计算均差表的一个算法可能是非常有效的，并且推荐它作为求插值多项式的最好方法，我们改变记号使得均差表有如下给出的元素：

$$
\begin{array}{cc|cccccc}
x_0 & c_{00} & c_{01} & c_{02} & c_{03} & \cdots & c_{0,n-1} & c_{0,n} \\
x_1 & c_{10} & c_{11} & c_{12} & c_{13} & \cdots & c_{1,n-1} & \\
x_2 & c_{20} & c_{21} & c_{22} & c_{23} & \iddots & & \\
\vdots & \vdots & \vdots & \vdots & \iddots & & & \\
\vdots & \vdots & \vdots & \iddots & & & & \\
x_{n-1} & c_{n-1,0} & c_{n-1,1} & & & & & \\
x_n & c_{n0} & & & & & &
\end{array}
$$

此外，竖线把数据(位于左端)与被计算的元素分隔开来，很显然我们有

$$c_{ij} = f[x_i, x_{i+1}, \cdots, x_{i+j}]$$

从(13)式的直接转换得到下列算法：

```
for j=1 to n do
  for i=0 to n-j do
      c_ij ← (c_{i+1,j-1} − c_{i,j-1})/(x_{i+j} − x_i)
  end do
end do
```

331

在这个算法中，数值 c_{i0}(它们是输入信息)是函数 f 在点 x_i 上的值，它们也是插值多项式在这些点上的值. 当然，插值多项式是

$$
\begin{aligned}
p(x) &= c_{00} + c_{01}(x-x_0) + c_{02}(x-x_0)(x-x_1) + \cdots \\
&\quad + c_{0n}(x-x_0)(x-x_1)\cdots(x-x_{n-1}) \\
&= \sum_{i=0}^{n} c_{0i} \prod_{j=0}^{i-1}(x-x_j)
\end{aligned}
$$

如果这个均差算法仅仅用来计算牛顿插值多项式的系数，那么我们能设计另一个占用更少计算机存储空间的算法，可以使用一个单一下标变量，记为 $d=[d_0, d_1, \cdots, d_n]$. 首先，把函数值 $f(x_0)$，$f(x_1)$，…，$f(x_n)$放入 d 中. 注意到 d_0 已经是牛顿插值多项式的第一项所求系数. 其次，我们计算均差的第一列，并把它们放在 d_1，d_2，…，d_n 的位置. 之后，d_0 仍然是其原有值，而 d_1 已经变为多项式的第二项所求系数，我们继续这种模式，仔细的储存新的均差在 d 向量的底部使得不打乱它的顶部次序，这样逐步得到最后的值. 算法如下：

```
for i=0 to n do
    d_i ← f(x_i)
end do
for j=1 to n do
```

```
    for i = n to j step -1 do
        d_i ← (d_i - d_{i-1})/(x_i - x_{i-j})
    end do
end do
```

根据算法的结论，向量 d 包含多项式的系数

$$p(x) = \sum_{i=0}^{n} d_i \prod_{j=0}^{i-1} (x - x_j)$$

在习题 6.2.10 中，将比较这个程序与 6.1 节中讨论的程序的有效性.

6.2.3 均差性质

332 我们用均差的一些良好性质来结束本节.

定理 2(均差排列定理) 均差是其自变量的对称函数. 因此，若$(z_0, z_1, \cdots, z_n)$是$(x_0, x_1, \cdots, x_n)$的一个排列，则

$$f[z_0, z_1, \cdots, z_n] = f[x_0, x_1, \cdots, x_n] \tag{15}$$

证明 (15)式左端的均差是 f 在点 $z_0, z_1, \cdots, z_n$ 上次数至多为 n 次的插值多项式中 x^n 的系数. (15)式右端的均差是 f 在点 $x_0, x_1, \cdots, x_n$ 上次数至多为 n 次的插值多项式中 x^n 的系数. 当然，这两个多项式是相等的. ■

定理 3(牛顿插值误差定理) 设 p 是函数 f 在 $n+1$ 个不同的结点 $x_0, x_1, \cdots, x_n$ 上次数至多为 n 次的插值多项式，若点 t 异于这些结点，则

$$f(t) - p(t) = f[x_0, x_1, \cdots, x_n, t] \prod_{j=0}^{n} (t - x_j) \tag{16}$$

证明 首先，设 q 是函数 f 在结点 $x_0, x_1, \cdots, x_n, t$ 上次数至多为 $n+1$ 次的插值多项式，我们知道 q 是由 p 添加一项得到的. 事实上

$$q(x) = p(x) + f[x_0, x_1, \cdots, x_n, t] \prod_{j=0}^{n} (x - x_j)$$

因为 $q(t)=f(t)$，所以我们立刻有(令 $x=t$)

$$f(t) = p(t) + f[x_0, x_1, \cdots, x_n, t] \prod_{j=0}^{n} (t - x_j)$$ ■

定理 4(导数和均差定理) 若 f 是$[a, b]$上 n 次连续可微函数，并且 $x_0, x_1, \cdots, x_n$ 是$[a, b]$中不同的点，则在(a, b)中存在一点 ξ，使得

$$f[x_0, x_1, \cdots, x_n] = \frac{1}{n!} f^{(n)}(\xi) \tag{17}$$

证明 首先，设 p 是函数 f 在结点 $x_0, x_1, \cdots, x_{n-1}$ 上次数至多为 $n-1$ 次的插值多项式. 根据 6.1 节定理 2 知，在(a, b)中存在一点 ξ，使得

$$f(x_n) - p(x_n) = \frac{1}{n!} f^{(n)}(\xi) \prod_{j=0}^{n-1} (x_n - x_j) \tag{18}$$

333 根据本节定理 3，我们有

$$f(x_n) - p(x_n) = f[x_0, x_1, \cdots, x_n] \prod_{j=0}^{n-1} (x_n - x_j) \tag{19}$$

比较(18)式和(19)式，可知(17)式成立. ■

6.2.4 Hermite-Genocchi 公式

许多情况需要一个称为 **Hermite-Genocchi 公式**的均差公式. 它指出均差 $f[x_0, x_1, \cdots, x_n]$是 $f^{(n)}(u_0x_0+\cdots+u_nx_n)$在 n 维单纯形上的积分，这个单纯形是下列$\mathbb{R}^{n+1}$中的集合 S_n

$$S_n = \{u = (u_0, u_1, \cdots, u_n) \in \mathbb{R}^{n+1}: u_i \geqslant 0, \sum_{i=0}^{n} u_i = 1\}$$

我们对 n 作数学归纳法证明 Hermite-Genocchi 公式. 对 $n=1$ 的情形，我们有

$$\begin{aligned} S_1 &= \{u = (u_0, u_1) \in \mathbb{R}^2: u_0 \geqslant 0, u_1 \geqslant 0, u_0 + u_1 = 1\} \\ &= \{(1-u_1, u_1): 0 \leqslant u_1 \leqslant 1\} \end{aligned}$$

因此，问题中的积分可写为

$$\begin{aligned} \int_{S_1} f'(u_0x_0 + u_1x_1)\mathrm{d}u &= \int_0^1 f'((1-u_1)x_0 + u_1x_1)\mathrm{d}u_1 \\ &= \int_0^1 f'(x_0 + u_1(x_1 - x_0))\mathrm{d}u_1 \\ &= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}u_1}[f(x_0 + u_1(x_1 - x_0))]\frac{\mathrm{d}u_1}{x_1 - x_0} \\ &= \frac{1}{x_1 - x_0} f(x_0 + u_1(x_1 - x_0))\Big|_{u_1=0}^{u_1=1} \\ &= \frac{1}{x_1 - x_0}\{f(x_1) - f(x_0)\} = f[x_0, x_1] \end{aligned}$$

根据归纳步骤，假设对 $n-1$ 的情形公式是正确的，那么对 n 的情形给予证明. 定义

$$I(x_0, x_1, \cdots, x_n) = \int_{S_n} f^{(n)}(u_0x_0 + \cdots + u_nx_n)\mathrm{d}u$$

因为 $\sum_{i=0}^{n} u_i = 1$，所以 $u_0 = 1 - \sum_{i=1}^{n} u_i$，因此

$$\begin{aligned} I(x_0, x_1, \cdots, x_n) &= \int_{S_n} f^{(n)}(x_0 + u_1(x_1 - x_0) + \cdots + u_n(x_n - x_0))\mathrm{d}u \\ &= \int_0^1 \int_0^{1-u_1} \cdots \int_0^{1-u_1-\cdots-u_{n-1}} f^{(n)}(x_0 + u_1(x_1 - x_0) \\ &\quad + \cdots + u_n(x_n - x_0))\mathrm{d}u_n \cdots \mathrm{d}u_2\mathrm{d}u_1 \end{aligned}$$

[334]

像 $n=1$ 的情形一样，可求出最内层的积分；即

$$\begin{aligned} &\int_0^{1-u_1-\cdots-u_{n-1}} \frac{\mathrm{d}}{\mathrm{d}u_n}[f^{(n-1)}(x_0 + u_1(x_1 - x_0) + \cdots + u_n(x_n - x_0))]\frac{\mathrm{d}u_n}{x_n - x_0} \\ &= \frac{1}{x_n - x_0}[f^{(n-1)}(x_0 + u_1(x_1 - x_0) + \cdots + u_n(x_n - x_0))]\Big|_{u_n=0}^{u_n=1-u_1-\cdots-u_{n-1}} \\ &= \frac{1}{x_n - x_0}\Big[f^{(n-1)}\Big(x_0 + \sum_{i=1}^{n-1} u_i(x_i - x_0) + \Big(1 - \sum_{i=1}^{n-1} u_i\Big)(x_n - x_0)\Big) \\ &\quad - f^{(n-1)}\Big(x_0 + \sum_{i=1}^{n-1} u_i(x_i - x_0)\Big)\Big] \end{aligned}$$

$$=\frac{1}{x_n-x_0}\left[f^{(n-1)}\left(x_n+\sum_{i=1}^{n-1}u_i(x_i-x_n)\right)-f^{(n-1)}\left(x_0+\sum_{i=1}^{n-1}u_i(x_i-x_0)\right)\right]$$

因此，积分 I 变为

$$\begin{aligned}&I(x_0,x_1,\cdots,x_n)\\&=\frac{1}{x_n-x_0}\int_0^1\int_0^{1-u_1}\cdots\int_0^{1-u_1-\cdots-u_{n-2}}\left[f^{(n-1)}\left(x_n+\sum_{i=1}^{n-1}u_i(x_i-x_n)\right)\right.\\&\quad\left.-f^{(n-1)}\left(x_0+\sum_{i=1}^{n-1}u_i(x_i-x_0)\right)\right]\mathrm{d}u_{n-1}\cdots\mathrm{d}u_2\mathrm{d}u_1\\&=\frac{1}{x_n-x_0}[I(x_n,x_1,\cdots,x_{n-1})-I(x_0,x_1,\cdots,x_{n-1})]\end{aligned}$$

根据归纳假设，上述最后一个表达式是

$$\frac{1}{x_n-x_0}\{f[x_n,x_1,\cdots,x_{n-1}]-f[x_0,x_1,\cdots,x_{n-1}]\}$$

根据均差的递归公式和均差在自变数置换下的不变性，上面的表达式正好是 $f[x_0, x_1, \cdots, x_n]$.

习题 6.2

[335]

1. 根据课本中概述的步骤证明(12)式.
2. 证明：若 f 连续，则在 $\mathbb{R}^{n+1}$ 的开集上 $f[x_0, x_1, \cdots, x_n]$ 也连续，其中向量 $(x_0, x_1, \cdots, x_n)$ 的分量不同.
3. 设 $f\in C^n[a, b]$，证明：若 $x_0\in(a, b)$ 并且 $x_1, x_2, \cdots, x_n$ 都收敛于 x_0，则 $f[x_0, x_1, \cdots, x_n]$ 收敛于 $f^{(n)}(x_0)/n!$.
4. 证明：若 f 是一个 k 次多项式，则对于 $n>k$，$f[x_0, x_1, \cdots, x_n]=0$.
5. 证明：若 p 是一个次数至多是 n 次的多项式，则
$$p(x)=\sum_{i=0}^{n}p[x_0,x_1,\cdots,x_i]\prod_{j=0}^{i-1}(x-x_j)$$
6. 证明：均差是全体函数上的线性映射；即证明等式
$$(\alpha f+\beta g)[x_0,x_1,\cdots,x_n]=\alpha f[x_0,x_1,\cdots,x_n]+\beta g[x_0,x_1,\cdots,x_n]$$
7. 如定理 4 中指出的那样，均差 $f[x_0, x_1]$ 类似于一阶导数．那么试问均差具有类似于 $(fg)'=f'g+fg'$ 的性质吗？
8. 在结点 $x_0, x_1, \cdots, x_n$ 上，利用 6.1 节中定义的函数 l_i，对任意的函数 f，证明
$$\sum_{i=0}^{n}f(x_i)\ell_i(x)=\sum_{i=0}^{n}f[x_0,x_1,\cdots,x_i]\prod_{j=0}^{i-1}(x-x_j)$$
9. (续)证明公式
$$f[x_0,x_1,\cdots,x_n]=\sum_{i=0}^{n}f(x_i)\prod_{\substack{j=0\\j\neq i}}^{n}(x_i-x_j)^{-1}$$
10. 比较均差算法与 6.1 节中用来计算牛顿插值多项式系数算法的效率.
11. 用矩阵理论中的克拉默法则证明
$$f[x_0,x_1,\cdots,x_n]=\begin{vmatrix}1&x_0&x_0^2&\cdots&x_0^{n-1}&f(x_0)\\1&x_1&x_1^2&\cdots&x_1^{n-1}&f(x_1)\\\vdots&\vdots&\vdots&&\vdots&\vdots\\1&x_n&x_n^2&\cdots&x_n^{n-1}&f(x_n)\end{vmatrix}\div\begin{vmatrix}1&x_0&x_0^2&\cdots&x_0^n\\1&x_1&x_1^2&\cdots&x_1^n\\\vdots&\vdots&\vdots&&\vdots\\1&x_n&x_n^2&\cdots&x_n^n\end{vmatrix}$$

12. 对于特殊的函数 $f(x)=x^m$，$m\in\mathrm{N}$，证明

$$f[x_0,x_1,\cdots,x_n]=\begin{cases}1 & 若\ n=m\\ 0 & 若\ n>m\end{cases}$$

13. 证明莱布尼茨公式

$$(fg)[x_0,x_1,\cdots,x_n]=\sum_{k=0}^{n}f[x_0,x_1,\cdots,x_k]g[x_k,x_{k+1},\cdots,x_n]$$

提示：参考上面的习题 6.2.7.

14. 把习题 6.2.9 中的公式改写为如下形式：

$$f[x_0,x_1,\cdots,x_n]=\sum_{i=0}^{n}a_i f(x_i)\qquad 其中\qquad a_i=\prod_{\substack{j=0\\ j\neq i}}^{n}(x_i-x_j)^{-1}$$

336

证明：若 x_i 排序如下：

$$x_0<x_1<x_2<\cdots<x_n$$

则 a_i 的符号交替变化.

15.（续）证明

$$\sum_{i=0}^{n}a_i x_i^n=1\quad 且\quad \sum_{i=0}^{n}a_i=\begin{cases}1\ 若\ n=0\\ 0\ 若\ n>0\end{cases}$$

16. 设 $f(x)=1/x$. 证明

$$f[x_0,x_1,\cdots,x_n]=(-1)^n\prod_{i=0}^{n}x_i^{-1}$$

17. 对下列表值求牛顿插值多项式.

x	1	3/2	0	2
$f(x)$	3	13/4	3	5/3

18. 证明：若 f 是多项式，则均差 $f[x_0,\ x_1,\ \cdots,\ x_n]$是变量 $x_0,\ x_1,\ \cdots,\ x_n$ 的多项式.

19. 证明：若 u 是 f 在点 $x_0,\ x_1,\ \cdots,\ x_{n-1}$ 上的任意插值函数，并且 v 是 f 在点 $x_1,\ x_2,\ \cdots,\ x_n$ 上的一个插值函数，则

$$[(x_n-x)u(x)+(x-x_0)v(x)]/(x_n-x_0)$$

是 f 在点 $x_0,\ x_1,\ \cdots,\ x_n$ 上的插值函数.

20.（续）考察数组

$$\begin{array}{cc|ccc} x_0 & y_0 & a_0 & b_0 & c_0\\ x_1 & y_1 & a_1 & b_1 & \\ x_2 & y_2 & a_2 & & \\ x_3 & y_3 & & & \end{array}$$

其中，对某些固定的 x，由下列公式计算 a_i，b_i 和 c_i.

$$a_i=[(x_{i+1}-x)y_i+(x-x_i)y_{i+1}]/(x_{i+1}-x_i)$$
$$b_i=[(x_{i+2}-x)a_i+(x-x_i)a_{i+1}]/(x_{i+2}-x_i)$$
$$c_i=[(x_{i+3}-x)b_i+(x-x_i)b_{i+1}]/(x_{i+3}-x_i)$$

证明：c_0 是三次插值多项式在 x 处的值.

21.（续）对任意的 n，推广上题中指出的计算 $p_n(x)$的算法．这种算法称为尼维尔算法.

22. 对下列表值求牛顿插值多项式.

x	0	1	2	7
y	51	3	1	201

[337] 23. 设 $p(x)=2-(x+1)+x(x+1)-2x(x+1)(x-1)$ 是下列表值中前4点上的插值多项式

x	-1	0	1	2	3
y	2	1	2	-7	10

试在 p 上添加一项，求出关于整个表值的一个插值多项式.

24. 对于下列表值写出牛顿插值多项式.

x	4	2	0	3
$f(x)$	63	11	7	28

25. 构造一个求解 $f(x)=0$ 迭代法如下：令 q_2 是下列表值的二次插值多项式，并且设 x_{n+1} 是 q_2 的最靠近 x_n 的零点：

$f(x_n)$	$f(x_{n-1})$	$f(x_{n-2})$
x_n	x_{n-1}	x_{n-2}

26. 证明：对于 $h>0$，有

$$f(x+2h)-2f(x+h)+f(x)=h^2 f''(\xi)$$

其中 ξ 在区间 $(x, x+2h)$ 内.

27. 证明：

$$m!f[0,1,\cdots,m]=\sum_{j=0}^{m}(-1)^{m-j}\binom{m}{j}f(j)$$

提示：用数学归纳法，在你的证明过程中，你必须停下来证明帕斯卡三角形中的基本恒等式：

$$\binom{m}{j-1}+\binom{m}{j}=\binom{m+1}{j}$$

28. 假设把 Hremite-Genocchi 公式表述为均差定义的形式，在此基础上，推导出递归公式(11).

29. 证明：若 $f^{(n)}$ 连续，则 $f[x_0, x_1, \cdots, x_n]$ 在 $\mathbb{R}^{n+1}$ 中处处连续，根据此结论推广习题 6.2.2 中的结果.

30. 在定理 4 中，ξ 是连续地依赖于 $x_0, x_1, \cdots, x_n$ 吗？对于 $f^{(n)}(\xi)$，回答相同的问题.

计算机习题 6.2

1. 在区间 $[-5, 5]$ 中对于函数 $f(x)=1/(1+x^2)$，求 $n=5$, 10 和 15 时的牛顿插值多项式 p_n. 用等距结点. 在每种情况下，都对 $[-5, 5]$ 中 30 个等距结点计算 $f(x)-p_n(x)$，以便观察 p_n 对 f 的发散程度.
2. 编写上面习题 6.2.25 中所述算法的程序并进行测试.

6.3 埃尔米特插值

[338] 术语**埃尔米特插值**指的是一个函数及其某些导数在一组结点上的插值.（后面会给出精确的定义.）当弄清楚这种类型的插值和比较简单类型的插值(其中没有导数的插值)之间的区别时，常称后者为**拉格朗日插值**.

6.3.1 基本概念

下面是埃尔米特插值的一个具有启发性和实用性的例子：我们要求函数 f 和它的导数 f' 在两个不同结点 x_0 和 x_1 上的一个次数最低的插值多项式. 这个多项式应该满足以下 4 个条件：

$$p(x_i)=f(x_i) \qquad p'(x_i)=f'(x_i) \qquad (i=0,1)$$

因为有 4 个条件，所以自然要在次数不超过 3 次的全体多项式组成的线性空间 Π_3 中求解，Π_3 中供我们选择的多项式有 4 个系数. 然而，我们不是把 $p(x)$ 写成 1，x，x^2，x^3 的组合形式，而是写为：

$$p(x) = a + b(x-x_0) + c(x-x_0)^2 + d(x-x_0)^2(x-x_1)$$

这种写法可以简化某些工作. 由此导出

$$p'(x) = b + 2c(x-x_0) + 2d(x-x_0)(x-x_1) + d(x-x_0)^2$$

现在 p 满足的 4 个条件可以写成下列形式

$$\begin{aligned} f(x_0) &= a \\ f'(x_0) &= b \\ f(x_1) &= a + bh + ch^2 \qquad (h = x_1 - x_0) \\ f'(x_1) &= b + 2ch + dh^2 \end{aligned}$$

显然，可以立刻得到 a 和 b. 把第 3 个等式中含 a 和 b 的项移到等式的左端，可以求出 c. 最后，从第 4 个等式可以确定 d. 因此，无论 $f(x_i)$ 和 $f'(x_i)$ 如何取值，总可以求得该问题的解.

一般来说，如果我们用一个多项式去插值一个函数及其某些导数的值，由于线性方程组(我们希望用它来计算多项式的系数)可能是奇异的，所以将会遇到一些困难. 一个简单的例子可以说明这一点.

例 1 求一多项式 p，使得：$p(0)=0$，$p(1)=1$，$p'(1/2)=2$.

解 因为有 3 个条件，我们试用一个二次多项式：

$$p(x) = a + bx + cx^2$$

由条件 $p(0)=0$ 可得 $a=0$. 另外两个条件导出

339

$$1 = p(1) = b + c$$

$$2 = p'(\frac{1}{2}) = b + c$$

因此，该问题没有二次多项式解. 我们注意到它的系数矩阵是奇异的. 对于同一个问题，如果我们试用一个三次多项式，

$$p(x) = a + bx + cx^2 + dx^3$$

就会发现这样的解存在，但不是唯一的. 如前所述，我们有 $a=0$，其他条件是

$$1 = b + c + d$$

$$2 = b + c + \frac{3}{4}d$$

这个方程组的解是 $d=-4$ 以及 $b+c=5$. ■

这种类型的一般问题显然会具有某些令人感兴趣的困难. 在称为**伯克霍夫插值**的专题里面，就奉献了大量的近期研究的文献. 想进一步研究该专题的读者可以参阅 Lorentz, Jetter, and Riemenschneider[1983].

我们现在要讨论具有唯一解的一大类插值问题. 这些问题人们一般称为**埃尔米特插值**. 在埃尔米特问题中，总假设给定导数 $p^{(j)}(x_i)$(在结点 x_i 上)，并且，$p^{(j-1)}(x_i)$，$p^{(j-2)}(x_i)$，…，$p'(x_i)$ 和 $p(x_i)$ 也是给定的. 我们用记号 k_i 表示在结点 x_i 上给定的 k_i 个插值条件. 注意到 k_i

会随着 i 的不同而变化. 设 x_0，x_1，…，x_n 是结点并且在 x_i 上给定的插值条件是：

$$p^{(j)}(x_i) = c_{ij} \qquad (0 \leqslant j \leqslant k_i - 1, 0 \leqslant i \leqslant n) \tag{1}$$

把多项式 p 上的插值条件总数记为 $m+1$，因此

$$m + 1 = k_0 + k_1 + \cdots + k_n \tag{2}$$

定理 1（埃尔米特插值定理） Π_m 中存在唯一的多项式 p 满足(1)式中的埃尔米特插值条件.

证明 因为在空间 Π_m 中寻求多项式 p，所以它有 $m+1$ 个系数. (1)式中对 p 强加的插值条件数也是 $m+1$. 从而，我们需要求解一个具有 $m+1$ 个未知量和 $m+1$ 个方程的方程组，并且希望保证它的系数矩阵非奇异. 要证明一个方阵 A 非奇异，只需证明齐次线性方程组 $Au=0$ 只有 0 解($u=0$)即可. 在下面讨论的插值问题中，**齐次问题**就是求出 $p \in \Pi_m$ 使得

340

$$p^{(j)}(x_i) = 0 \qquad (0 \leqslant j \leqslant k_i - 1, 0 \leqslant i \leqslant n)$$

$x_i(0 \leqslant i \leqslant n)$是这个多项式的 k_i 重零点，因此它一定是下列多项式 q 的倍数：

$$q(x) = \prod_{i=0}^{n} (x - x_i)^{k_i}$$

然而，可以看出 q 的次数是

$$m + 1 = \sum_{i=0}^{n} k_i$$

而 p 的次数至多是 m. 因此得到 $p=q=0$. ■

例 2 当只有一个结点时，埃尔米特插值是什么形式?

解 在这种情况下，我们需要一个 k 次多项式 p，使得

$$p^{(j)}(x_0) = c_{0j} \qquad (0 \leqslant j \leqslant k)$$

其解是泰勒多项式

$$p(x) = c_{00} + c_{01}(x - x_0) + \frac{c_{02}}{2!}(x - x_0)^2 + \cdots + \frac{c_{0k}}{k!}(x - x_0)^k$$ ■

6.3.2 牛顿均差方法

现在我们解释如何推广牛顿均差方法用来求解埃尔米特插值问题. 先从一种简单情形开始，寻求一个取下列给定值的二次多项式：

$$p(x_0) = c_{00} \qquad p'(x_0) = c_{01} \qquad p(x_1) = c_{10} \tag{3}$$

我们把均差表写成如下形式：

x_0	c_{00}	c_{01}	?
x_0	c_{00}	?	
x_1	c_{10}		

表中的问号表示还没有计算出的元. 由于两个条件都在 x_0 用到 p，我们看到 x_0 在自变量列中出现了两次. 进一步注意到 $p'(x_0)$的给定值位于一阶均差所在的列. 这与下列等式一致：

$$\lim_{x \to x_0} f[x_0, x] = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0)$$

这个等式证明了下列定义是有道理的

$$f[x_0,x_0] \equiv f'(x_0)$$

341

均差表中其他元可由平常的方法计算．当结点重复时可以预计到的困难仅出现在 c_{01} 处，而 c_{01} 的值已经由数据提供，所以不需要计算 c_{01}．用问号表示的元可以用常规方法计算：

$$p[x_0,x_1] = \frac{p(x_1)-p(x_0)}{x_1-x_0} = \frac{c_{10}-c_{00}}{x_1-x_0} \tag{4}$$

和

$$p[x_0,x_0,x_1] = \frac{p[x_0,x_1]-p[x_0,x_0]}{x_1-x_0} = \frac{c_{10}-c_{00}}{(x_1-x_0)^2} - \frac{c_{01}}{x_1-x_0} \tag{5}$$

插值多项式可以写成常见的形式

$$p(x) = p(x_0) + p[x_0,x_0](x-x_0) + p[x_0,x_0,x_1](x-x_0)^2 \tag{6}$$

这个多项式是(3)式问题的解，其直接的证明由习题 6.3.5 给出．

回到本节开始时候的例题，我们已得到插值多项式

$$p(x) = f(x_0) + f'(x_0)(x-x_0) + f[x_0,x_0,x_1](x-x_0)^2 + f[x_0,x_0,x_1,x_1](x-x_0)^2(x-x_1)$$

它直接来自于下列均差表

x_0	$f(x_0)$	$f'(x_0)$	$f[x_0,\ x_0,\ x_1]$	$f[x_0,\ x_0,\ x_1,\ x_1]$
x_0	$f(x_0)$	$f[x_0,\ x_1]$	$f[x_0,\ x_1,\ x_1]$	
x_1	$f(x_1)$	$f'(x_1)$		
x_1	$f(x_1)$			

所有自变量相同的均差与 6.2 节定理 4 中的一致．该定理保证存在一点 ξ，使得

$$f[x_0,x_1,\cdots,x_k] = \frac{1}{k!}f^{(k)}(\xi)$$

这里必须假设 $f^{(k)}$ 存在并且在包含结点 x_0，x_1，…，x_k 的最小区间内连续．点 ξ 也位于这个小区间内．如果该区间的长度收缩到 0，我们得到极限

$$f[x_0,x_0,\cdots,x_0] = \frac{1}{k!}f^{(k)}(x_0) \tag{7}$$

注意，当 $k \geqslant 2$ 时，我们一定要留神包含因子 $1/k$！

342

例 3 用推广的牛顿均差算法确定一个多项式，使得它取值如下：

$$p(1)=2 \qquad p'(1)=3 \qquad p(2)=6 \qquad p'(2)=7 \qquad p''(2)=8$$

解 我们把数据排列成下面的均差表，并用问号表示待计算的量．而在右端的表中给出最后的计算结果．

1	2	3	?	?	?
1	2	?	?	?	
2	6	7	4		
2	6				
2	6				

1	2	3	1	2	−1
1	2	4	3	1	
2	6	7	4		
2	6				
2	6				

注意，表中第 3 行的第 2 个均差 4 是根据(7)式中 $k=2$ 的情形得到的．算出全部均差以后，表中第一行的数字(除结点之外)就是插值多项式的系数：

$$p(x)=2+3(x-1)+(x-1)^2+2(x-1)^2(x-2)-(x-1)^2(x-2)^2 \quad \blacksquare$$

6.3.3 拉格朗日型

原则上来说，对埃尔米特插值也可以讨论像拉格朗日插值公式那样的公式．我们将介绍一个适合于重要特殊情况的这样的公式．如前所述，设 x_0，x_1，…，x_n 是结点，并且假设在每个结点上的函数值和它的一阶导数是给定的．我们要寻找的多项式 p 必须满足下列等式：

$$p(x_i)=c_{i0} \qquad p'(x_i)=c_{i1} \qquad (0\leqslant i\leqslant n) \tag{8}$$

与拉格朗日公式相类似，我们有

$$p(x)=\sum_{i=0}^{n}c_{i0}A_i(x)+\sum_{i=0}^{n}c_{i1}B_i(x) \tag{9}$$

其中 A_i 和 B_i 是具有某些特殊性质的多项式．稍微想想就会发现下列性质完全符合我们的要求：

$$\begin{cases}A_i(x_j)=\delta_{ij}\\A'_i(x_j)=0\end{cases} \qquad \begin{cases}B_i(x_j)=0\\B'_i(x_j)=\delta_{ij}\end{cases}$$

当然，如果 A_i 和 B_i 具有上述性质，很简单的就可以证明(9)式中给出的多项式 p 满足(8)式中的插值性质．这一点可借助于下列函数来验证：

343

$$\ell_i(x)=\prod_{\substack{j=0\\j\neq i}}^{n}\frac{x-x_j}{x_i-x_j} \qquad (0\leqslant i\leqslant n)$$

A_i 和 B_i 可定义如下：

$$\begin{cases}A_i(x)=[1-2(x-x_i)\ell'_i(x_i)]\ell_i^2(x) & (0\leqslant i\leqslant n)\\B_i(x)=(x-x_i)\ell_i^2(x) & (0\leqslant i\leqslant n)\end{cases}$$

注意到每一个 ℓ_i 都是 n 次多项式，因此 A_i 和 B_i 的次数是 $2n+1$．从而，p 的次数至多是 $2n+1$ 次．因为(8)式中要求多项式 p 满足 $2n+2$ 个条件，所以 p 的次数恰好符合要求．

例 1 中的拉格朗日型插值多项式是：

$$p(x)=f(x_0)A_0(x)+f(x_1)A_1(x)+f'(x_0)B_0(x)+f'(x_1)B_1(x)$$

其中

$$A_0(x)=[1-2(x-x_0)\ell'_0(x_0)]\ell_0^2(x)$$
$$A_1(x)=[1-2(x-x_1)\ell'_1(x_1)]\ell_1^2(x)$$
$$B_0(x)=(x-x_0)\ell_0^2(x)$$
$$B_1(x)=(x-x_1)\ell_1^2(x)$$

并且

$$\ell_0(x)=\frac{x-x_1}{x_0-x_1}$$

$$\ell_1(x)=\frac{x-x_0}{x_1-x_0}$$

$$\ell'_0(x) = \frac{1}{x_0 - x_1}$$

$$\ell'_1(x) = \frac{1}{x_1 - x_0}$$

下面的定理给出一个这种类型的埃尔米特插值误差公式

定理 2(埃尔米特插值误差估计定理) 设 x_0, x_1, …, x_n 是区间$[a, b]$中不同的结点, 并且 $f \in C^{2n+2}[a, b]$. 若次数至多为 $2n+1$ 次的多项式 p 使得

$$p(x_i) = f(x_i) \qquad p'(x_i) = f'(x_i) \qquad (0 \leqslant i \leqslant n)$$

则对$[a, b]$中的每一点 x, 都有(a, b)中的一点 ξ 使得

$$f(x) - p(x) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \prod_{i=0}^{n} (x - x_i)^2$$

344

证明 如果 x 是一个结点, 公式显然是正确的. 现取定 x(不是结点), 定义 w 和 ϕ 如下:

$$w(t) = \prod_{i=0}^{n} (t - x_i)^2 \qquad \phi = f - p - \lambda w$$

其中选取 λ 满足 $\phi(x)=0$. 注意到 ϕ 在$[a, b]$中至少有 $n+2$ 个零点; 即 x, x_0, x_1, …, x_n. 根据罗尔定理, ϕ' 至少有 $n+1$ 个零点, 它们不同于前面已列出的点. 此外, ϕ' 在每个结点上都是零. 因此, ϕ' 在$[a, b]$内至少有 $2n+2$ 个零点. 根据罗尔定理, ϕ'' 在(a, b)内至少有 $2n+1$ 个零点. 重复上面的讨论, 我们得到 $\phi^{(2n+2)}$ 在(a, b)内有一个零点 ξ, 从而

$$0 = \phi^{(2n+2)}(\xi) = f^{(2n+2)}(\xi) - p^{(2n+2)}(\xi) - \lambda w^{(2n+2)}(\xi)$$

因为 p 至多是 $2n+1$ 次多项式, 所以 $p^{(2n+2)}=0$. 由于 $w(t)$ 的首项是 t^{2n+2}, 从而有 $w^{(2n+2)}=(2n+2)!$. 利用这个信息以及 $\phi^{(2n+2)}$ 等式中的值 $\lambda=[f(x)-p(x)]/w(x)$, 我们有

$$0 = f^{(2n+2)}(\xi) - [f(x) - p(x)](2n+2)!/w(x)$$

这是所证结论的另一种表达形式. ■

6.3.4 带重复结点的均差

本节的剩余部分, 我们专门讨论自变量可以重复的均差. 有多种方法可以开始我们的讨论. 例如, 可以用递归定义(Braess[1984]), 或者用行列式引入的定义(Schumaker[1981]), 也可以简单的推广不同自变量均差 $f[x_0, x_1, \cdots, x_n]$ 的定义. 我们采用最后一种方式, 它由 Conte and de Boor[1980]给出.

回顾前面所学, 如果 x_0, x_1, …, x_n 是不同的点, 则 f 在结点 x_0, x_1, …, x_n 上的插值多项式在 Π_n 中, 多项式中 x^n 的系数为 6.2 节中定义的 $f[x_0, x_1, \cdots, x_n]$. 当这些点中包含重复结点时, 我们给出插值的定义如下:

定义 1(带重复结点插值的定义) 若对于在结点表 x_0, x_1, …, x_n 中重复出现 k 次或者更多次的每一点 ξ, 都有 $f^{(k-1)}(\xi)=0$, 则我们称 f 在点 x_0, x_1, …, x_n 上插值零.

例如, 我们称 f 在点 1, 3, 8, 1, 13, 1, 8 上插值零, 如果

$$f(1) = f(3) = f(8) = f'(1) = f(13) = f''(1) = f'(8) = 0$$

很容易证明多项式 $p(x)$ 在点 x_0, x_1, …, x_n 上插值零的充分必要条件是 $p(x)$ 包含下列因式:

345

$$q(x)=\prod_{j=0}^{n}(x-x_j)$$

如果两个函数 f 和 g 使得 $f-g$ 在点 x_0，x_1，…，x_n 上插值零，我们称 f 在点 x_0，x_1，…，x_n 上插值 g（或者 g 在点 x_0，x_1，…，x_n 上插值 f）.

利用刚才说明的这些术语，我们可以把本节中已证明的定理重新叙述如下：

定理 3（多项式插值的唯一性定理） *设 x_0，x_1，…，x_m 是一串结点，其中相同的点最多重复出现 k 次. 设 f 属于包含这些结点的某一区间上的函数类 C^{k-1}，则在 Π_m 中存在唯一一个多项式在这些结点上插值 f.*

一般情况下，f 在结点 x_0，x_1，…，x_n 上的插值多项式在 Π_n 中，并且 x^n 的系数定义为 $f[x_0, x_1, \cdots, x_n]$. 如果点 ξ 在结点表中重复出现 k 次，那么该定义就需要导数 $f^{(k-1)}(\xi)$的存在. 否则，均差就不存在(即没有定义). 依据这个定义，我们可以证明**一般牛顿插值公式**的正确性

$$p(x)=\sum_{j=0}^{n}f[x_0,x_1,\cdots,x_j]\prod_{i=0}^{j-1}(x-x_i) \tag{10}$$

记住，由定义 $\prod_{i=0}^{-1}(x-x_j)=1$.

定理 4（一般牛顿插值多项式定理） *若 f 充分可微，使得等式(10)中的均差存在，则(10)式给出 Π_n 中的多项式 p 在点 x_0，x_1，…，x_n 上插值 f.*

证明 对 n 作数学归纳法. 对于 $n=0$，$f[x_0]$是 Π_0 中多项式在点 x_0 上插值 f. 这一点显然正确.

现在定义如下的多项式 q

$$q(x)=\sum_{j=0}^{n-1}f[x_0,x_1,\cdots,x_j]\prod_{i=0}^{j-1}(x-x_i)$$

在点 x_0，x_1，…，x_{n-1}上插值 f. 设 p 是 Π_n 中的多项式在点 x_0，x_1，…，x_n 上插值 f. 前面的定理保证了 p 的存在性和唯一性. 根据均差的定义，p 中 x^n 的系数是 $f[x_0, x_1, \cdots, x_n]$. 因此多项式

$$p(x)-f[x_0,x_1,\cdots,x_n]\prod_{i=0}^{n-1}(x-x_i)$$

的次数至多是 $n-1$ 次，根据习题 6.3.8、6.3.9 和 6.3.11 知，该多项式在点 x_0，x_1，…，x_{n-1}上插值 f. 由 q 的唯一性，一定有

346

$$p(x)-f[x_0,x_1,\cdots,x_n]\prod_{i=0}^{n-1}(x-x_i)=q(x)$$

因此，我们有

$$\begin{aligned}p(x)&=q(x)+f[x_0,x_1,\cdots,x_n]\prod_{i=0}^{n-1}(x-x_i)\\&=\sum_{j=0}^{n}f[x_0,x_1,\cdots,x_j]\prod_{i=0}^{j-1}(x-x_i)\end{aligned}$$

■

最后要解决的问题是一般形式的均差是否满足递归关系

$$f[x_0,x_1,\cdots,x_n]=\{f[x_1,x_2,\cdots,x_n]-f[x_0,x_1,\cdots,x_{n-1}]\}/(x_n-x_0)$$

假设 $x_0\leqslant x_1\leqslant\cdots\leqslant x_n$. 因为一般均差是关于自变量的对称函数，所以该假设不影响均差的一般性.（一般情况下它也是正确的，在 6.2 节中给出了这个事实的证明.）递归关系在 $x_n=x_0$ 时显然不成立，不过这意味着 $x_0=x_1=\cdots=x_n$，此时我们恰好用到公式

$$f[x_0,x_0,\cdots,x_0]=\frac{1}{n!}f^{(n)}(x_0)$$

除此以外，其他情况下递归公式都成立. 下面我们正式地给出证明.

定理 5（均差递归公式定理） 设 $x_0\leqslant x_1\leqslant\cdots\leqslant x_n$，则均差服从下列递归公式：

$$f[x_0,x_1,\cdots,x_n]=\begin{cases}\dfrac{f[x_1,x_2,\cdots,x_n]-f[x_0,x_1,\cdots,x_{n-1}]}{x_n-x_0} & 若\ x_n\neq x_0\\ f^{(n)}(x_0)/n! & 若\ x_n=x_0\end{cases}\tag{11}$$

证明 对 n 作数学归纳法. Π_1 中在点 x_0 和 x_1 上插值 f 的多项式是

$$p(x)=\begin{cases}(x-x_0)[f(x_1)-f(x_0)]/(x-x_0)+f(x_0) & (x_0\neq x_1)\\ f'(x_0)(x_1-x_0)+f(x_0) & (x_0=x_1)\end{cases}$$

在这个多项式中，x 的系数是

$$f[x_0,x_1]=\begin{cases}\{f[x_1]-f[x_0]\}/(x_1-x_0) & (x_0\neq x_1)\\ f'(x_0) & (x_0=x_1)\end{cases}$$

这与(11)式中结果一致.

现在假设(11)式对从 1 到 $m-1$ 的所有整数 n 都成立. 设 $x_0\leqslant x_1\leqslant\cdots\leqslant x_m$ 是一串结点，考虑 Π_m 中在这些结点上插值 f 的多项式 p. 如果 $x_m=x_0$，那么所有结点 x_0，x_1，$\cdots$，x_m 都相同. 此时，由习题 6.3.7 知 p 是 f 在点 x_0 上的泰勒多项式：

347

$$p(x)=\sum_{k=0}^{m}\frac{1}{k!}f^{(k)}(x_0)(x-x_0)^k$$

$p(x)$ 中 x^m 的系数是 $f^{(m)}(x_0)/m!$，这表明当 $n=m$ 且 $x_m=x_0$ 时(11)式是正确的. 而 6.2 节定理 1 的证明方法说明了另一种情况也是正确的. ■

习题 6.3

1. 用推广的牛顿均差方法求一个满足下列表值的二次多项式：

x	0	1	2
$p(x)$	2	-4	44
$p'(x)$	-9	4	

2.（续）求一个满足上题中表值，并且满足 $p(3)=2$ 的五次多项式 p. 提示：在上题求出的多项式中添加适当的项.

3. 试求满足下列表值的次数最低的多项式 p 的公式：

$$p(x_i)=y_i\qquad p'(x_i)=0\qquad (0\leqslant i\leqslant n)$$

4. 如果插值问题

$$p(x_i)=c_{i0}\qquad p''(x_i)=c_{i2}\qquad (i=0,1)$$

有一个三次多项式的解（对任意的 c_{ij}），试问对结点 x_0 和 x_1 应设置什么条件？

5. 利用(4)式和(5)式中给出的均差，证明等式(6)给出的多项式满足(3)式中的条件.
6. 证明函数 A_i 和 B_i 所具有的性质.
7. 证明泰勒多项式

$$p(x)=\sum_{j=0}^{k-1}\frac{1}{j!}f^{(j)}(x_0)(x-x_0)^j$$

在点 x_0，x_0，…，x_0(k 重)上插值 f.

8. 证明一个多项式在点 x_0，x_1，…，x_n(允许重复)上插值零的充分必要条件是它包含因子 $\prod_{j=0}^{n}(x-x_j)$.
9. 证明：若 f 在点 x_0，x_1，…，x_n 上插值 g，并且 h 在这些点上插值零，则 $f\pm ch$ 在这些点上插值 g.
10. 给定结点 x_0，x_1，…，x_n. 证明在这些点上插值零的函数集合构成一个代数；即对加法运算，乘法运算以及数乘运算是封闭的.
11. 证明：若 f 在结点 x_0，x_1，…，x_n 上插值零，则它在结点 x_0，x_1，…，x_{n-1} 上也插值零.
12. 设 $x_0<x_1<\cdots<x_n$，并且 f 连续可微. 证明：

348

$$\frac{\partial}{\partial x_i}f[x_0,x_1,\cdots,x_n]=f[x_0,x_1,\cdots,x_i,x_i,x_{i+1},\cdots,x_n]$$

13. 当 $n=2$ 时，写出(9)式中出现的函数 A_i 和 B_i 的具体形式，并化简这些函数.
14. 推导出 $\ell'_i(x)$的公式.
15.(利用罗尔定理)设 $f\in C^n[\alpha,\beta]$. 假设 f 在点 α 有 m 重零点而在点 β 有 k 重根，其中 $m\geqslant1$，$k\geqslant1$ 且 $m+k-1=n$. 证明 $f^{(n)}$ 在(α，β)中至少有一个零点.
16. 考虑多项式

$$p(t)=b-(b-a)\left[3\left(\frac{b-t}{b-a}\right)^2-2\left(\frac{b-t}{b-a}\right)^3\right]$$

证明：$|p'(t)|\leqslant p'((a+b)/2)=3/2$，$p(a)=a$，$p(b)=b$，$p'(a)=0$，以及 $p'(b)=0$.

6.4 样条插值

样条函数是由一些具有某些连续性条件的子区间上的分段多项式构成. 给定 $n+1$ 个点 t_0，t_1，…，t_n 并且满足 $t_0<t_1<\cdots<t_n$. 这些点称为**结点**. 又假如指定一个整数 $k\geqslant0$. 具有结点 t_0，t_1，…，t_n 的一个 **k 次样条函数**是指满足下列条件的函数 S：

1. 在每一个区间$[t_{i-1}, t_i)$上，S 是一个次数$\leqslant k$ 的多项式.
2. 在$[t_0, t_n]$上 S 有($k-1$)阶连续导数.

因此，S 是一个次数至多是 k 次的分段多项式，并且具有直到 $k-1$ 阶的连续导数.

0 次样条函数是分段常值函数，一个零次样条函数可以直接写成下列形式：

$$S(x)=\begin{cases}S_0(x)=c_0 & x\in[t_0,t_1)\\ S_1(x)=c_1 & x\in[t_1,t_2)\\ \quad\vdots & \quad\vdots\\ S_{n-1}(x)=c_{n-1} & x\in[t_{n-1},t_n]\end{cases}$$

区间$[t_{i-1}, t_i)$互不相交，因此函数在每个结点上的定义是确定的. 图 6-3 给出了具有 6 个结点的一个 0 次样条函数的图像. 图 6-4 给出了具有 9 个结点的一次样条函数的图像. 这样的函数可以直接定义为：

$$S(x)=\begin{cases}S_0(x)=a_0x+b_0 & x\in[t_0,t_1)\\ S_1(x)=a_1x+b_1 & x\in[t_1,t_2)\\ \vdots & \vdots\\ S_{n-1}(x)=a_{n-1}x+b_{n-1} & x\in[t_{n-1},t_n]\end{cases}$$

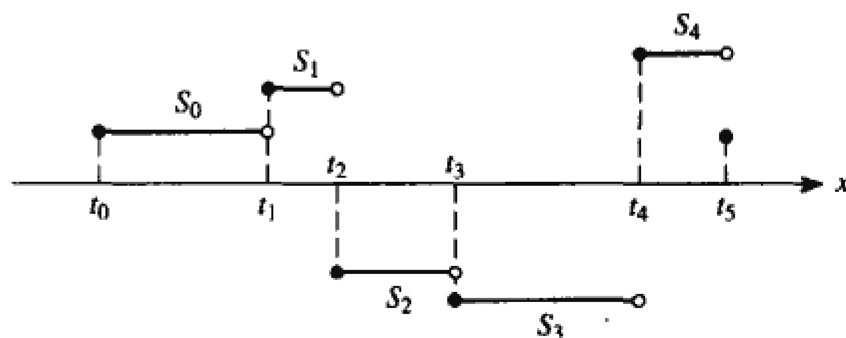


图 6-3　0 次样条函数

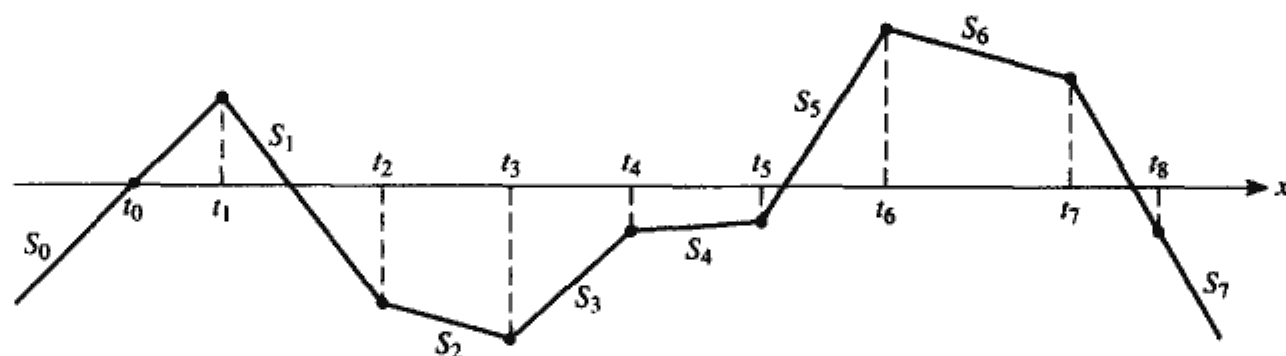


图 6-4　一次样条函数

如果给定所有的结点 t_i 及系数 a_i，b_i，那么通过首先确定包含 x 的子区间$[t_i,\ t_{i+1})$便可得到 S 在点 x 的值. 样条函数还可以定义在整个实轴上. 为方便起见，在区间$(-\infty,\ t_1)$上用表达式 a_0x+b_0，在区间$[t_{n-1},\ \infty)$上用表达式 $a_{n-1}x+b_{n-1}$. 因为函数 S 是连续的，所以分段多项式在结点相匹配，即 $S_i(t_{i+1})=S_{i+1}(t_{i+1})$. 下面是一个计算一次样条 $S(x)$的伪代码：

```
input (t_i), (a_i), (b_i), x, n
for i=1 to n-1 do
  if x<t_i then
    S(x)=a_{i-1}x+b_{i-1}
    output S(x)
    exit loop
  end if
end do
S(x)=a_{n-1}x+b_{n-1}
output S(x)
```

6.4.1　三次样条

因为**三次样条**($k=3$)有广泛的实际应用. 所以我们要十分详尽地讨论它的理论和构造. 我们假设给定下面的表值

x	t_0	t_1	$\cdots$	t_n
y	y_0	y_1	$\cdots$	y_n

(1)

349～350 并且对这个表值插值可以构造一个 3 次样条 S. 在每一个区间$[t_0, t_1]$, $[t_1, t_2]$, …, $[t_{n-1}, t_n]$上，S 都是不同的三次多项式. 我们把在$[t_i, t_{i+1}]$上表示 S 的多项式记为 S_i，从而，

$$S(x) = \begin{cases} S_0(x) & x \in [t_0, t_1] \\ S_1(x) & x \in [t_1, t_2] \\ \vdots & \vdots \\ S_{n-1}(x) & x \in [t_{n-1}, t_n] \end{cases} \tag{2}$$

多项式 S_{i-1}和 S_i 在点 t_i 上有相同的插值，所以

$$S_{i-1}(t_i) = y_i = S_i(t_i) \qquad (1 \leqslant i \leqslant n-1) \tag{3}$$

因此，S 是自动地连续的. 进一步，我们假设 S'和 S''也是连续的，在三次样条函数的求导中将会用到这些条件.

S，S'和 S''的连续性足够确定三次样条了吗？因为有 n 个三次多项式，每个多项式有 4 个系数，所以在这个分段三次多项式中有 $4n$ 个系数. 在每个子区间$[t_i, t_{i+1}]$上有两个插值条件：$S(t_i)=y_i$ 和 $S(t_{i+1})=y_{i+1}$，这就给出了 $2n$ 个条件. S 的连续性不再给出额外的条件，因为在插值条件中已使用过了. 在每一个内结点上，S'的连续性确定了一个条件 $S'_{i-1}(t_i)=S'_i(t_i)$，总共有 $n-1$ 个额外的条件. 类似地，S''的连续性也可以给出另外的 $n-1$ 个条件. 因此，总共有确定 $4n$ 个系数的 $4n-2$ 个条件. 剩余的自由度是 2，我们有合理使用这些自由度的各种方法.

现在我们推导出区间$[t_i, t_{i+1}]$上 $S_i(x)$的表达式. 首先，我们定义一组数 $z_i=S''(t_i)$. 因为 S''在每个内结点上连续，所以显然，对于 $0 \leqslant i \leqslant n$，存在 z_i 且满足

$$\lim_{x \downarrow t_i} S''(x) = z_i = \lim_{x \uparrow t_i} S''(x) \qquad (1 \leqslant i \leqslant n-1) \tag{4}$$

由于 S_i 是$[t_i, t_{i+1}]$上的三次多项式，因此 S''_i 是满足 $S''_i(t_i)=z_i$ 和 $S''_i(t_{i+1})=z_{i+1}$ 的线性函数，所以 S''_i 也是 z_i 和 z_{i+1}之间的直线：

$$S''_i(x) = \frac{z_i}{h_i}(t_{i+1} - x) + \frac{z_{i+1}}{h_i}(x - t_i) \tag{5}$$

其中 $h_i \equiv t_{i+1} - t_i$. 把这个函数积分两次，其结果是 S_i：

$$S_i(x) = \frac{z_i}{6h_i}(t_{i+1} - x)^3 + \frac{z_{i+1}}{6h_i}(x - t_i)^3 + C(x - t_i) + D(t_{i+1} - x) \tag{6}$$

其中 C 和 D 是积分常数. （验证：把(6)式微分两次可以得到(5)式.）将插值条件 $S_i(t_i)=y_i$ 和 $S_i(t_{i+1})=y_{i+1}$作用在 S_i 上就可以确定 C 和 D，其结果是

351

$$S_i(x) = \frac{z_i}{6h_i}(t_{i+1} - x)^3 + \frac{z_{i+1}}{6h_i}(x - t_i)^3 + \left(\frac{y_{i-1}}{h_i} - \frac{z_{i+1}h_i}{6}\right)(x - t_i) + \left(\frac{y_i}{h_i} - \frac{z_i h_i}{6}\right)(t_{i+1} - x) \tag{7}$$

容易证明(7)式是正确的. 只要简单地令 $x=t_i$ 及 $x=t_{i+1}$，就可以看出插值条件成立. 一旦确定了 z_0，z_1，…，z_n 的值以后，可以用(2)式和(7)式算出 $S(x)$在区间$[t_0, t_n]$内任意一点 x 处的值.

我们用S'的连续性条件来确定z_1，z_2，…，z_{n-1}．在内结点t_i上，一定有$S'_{i-1}(t_i)=S'_i(t_i)$．对(7)式求微分可得到$S'_i(x)$．然后作替换$x=t_i$并化简得

$$S'_i(t_i)=-\frac{h_i}{3}z_i-\frac{h_i}{6}z_{i+1}-\frac{y_i}{h_i}+\frac{y_{i+1}}{h_i} \tag{8}$$

同理，可由(7)式得到S'_{i-1}①，我们有

$$S'_{i-1}(t_i)=\frac{h_{i-1}}{6}z_{i-1}+\frac{h_{i-1}}{3}z_i-\frac{y_{i-1}}{h_{i-1}}+\frac{y_i}{h_{i-1}} \tag{9}$$

当(8)式和(9)式的右端项建立一个等式时，其结果可写成

$$h_{i-1}z_{i-1}+2(h_i+h_{i-1})z_i+h_iz_{i+1}=\frac{6}{h_i}(y_{i+1}-y_i)-\frac{6}{h_{i-1}}(y_i-y_{i-1}) \tag{10}$$

这个等式仅仅对$i=1,2,\cdots,n-1$成立．(为什么?)从而它给出了一个含有$n+1$个未知量z_0，z_1，…，z_n的$n-1$个方程的线性方程组．我们可以任意选择z_0和z_n来求解刚才所得的线性方程组并得到z_1，z_2，…，z_{n-1}．一个很好的选择是$z_0=z_n=0$，由此得到的样条函数称为**自然三次样条**．(后面将要证明的定理将为样条函数的这种选择提供依据．)

对于$1\leqslant i\leqslant n-1$，以及$z_0=z_n=0$，线性方程组(10)是对称的、三对角的和对角占优的，它具有下列形式

$$\begin{bmatrix} u_1 & h_1 & & & & \\ h_1 & u_2 & h_2 & & & \\ & h_2 & u_3 & h_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & h_{n-3} & u_{n-2} & h_{n-2} \\ & & & & h_{n-2} & u_{n-1} \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_{n-2} \\ z_{n-1} \end{bmatrix}=\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_{n-2} \\ v_{n-1} \end{bmatrix}$$

其中

$$h_i=t_{i+1}-t_i$$
$$u_i=2(h_i+h_{i-1})$$
$$b_i=\frac{6}{h_i}(y_{i+1}-y_i)$$
$$v_i=b_i-b_{i-1}$$

它可由下列特殊算法(不用行尺度主元的高斯消元法)求解：　352

```
input n, (t_i), (y_i)
for i = 0 to n-1 do
    h_i ← t_{i+1} - t_i
    b_i ← 6(y_{i+1} - y_i)/h_i
end do
u_1 ← 2(h_0 + h_1)
v_1 ← b_1 - b_0
```

---

① 原文误为S_{i-1}．——译者注

```
for i = 2 to n-1 do
    u_i ← 2(h_i + h_{i-1}) - h_{i-1}^2 / u_{i-1}
    v_i ← b_i - b_{i-1} - h_{i-1} v_{i-1} / u_{i-1}
end do
z_n ← 0
for i = n-1 to 1 step -1 do
    z_i ← (v_i - h_i z_{i+1}) / u_i
end do
z_0 ← 0
output (z_i)
```

在上述算法的基础上很容易编写出一个子程序或过程．它输入结点数组(t_i)和对应的函数值数组(y_i)，运行后给出数组(z_i)的值．

因为在这个算法中用到u_i做除法，所以需要证明$u_i \neq 0$．我们用数学归纳法证明$u_i > h_i > 0$．对于$i=1$，由于$u_1 = 2(h_0 + h_1)$，结论显然成立．假设$u_{i-1} > h_{i-1}$，因为

$$u_i = 2(h_i + h_{i-1}) - \frac{h_{i-1}^2}{u_{i-1}} > 2(h_i + h_{i-1}) - h_{i-1} > h_i = t_{i+1} - t_i > 0$$

因此$u_i > h_i$．

当确定系数z_0，z_1，…，z_n以后，三次样条函数(2)式的任意值都可通过(7)式来计算．任意给定x，首先需要确定下列区间

$$(-\infty, t_1), [t_1, t_2), \cdots, [t_{n-2}, t_{n-1}), [t_{n-1}, \infty)$$

中包含x的区间．为方便起见，我们约定，S_0不仅定义在$[t_0, t_1]$上，而且也定义在$(-\infty, t_0)$上．同样地，S_{n-1}不仅定义在$[t_{n-1}, t_n]$上而且也定义在(t_n, ∞)上．我们按顺序检验数组

$$x - t_{n-1}, x - t_{n-2}, \cdots, x - t_1$$

是否存在非负数来确定包含x的区间．如果其中一个数项非负，我们选取第一个非负数$x - t_i$，那$x - t_i \geqslant 0$但$x - t_{i+1} < 0$，因此$t_i \leqslant x < t_{i+1}$．如果所有检验数项都为负的，那么$x \in (-\infty, t_1]$．根据这样确定的指标$i$，用(7)式可以计算出所要求的多项式$S_i$在给定的点$x$的值．然而，利用习题 6.4.4，我们可以把(7)式改写成下列更有效的嵌套形式：

353

$$S_i(x) = y_i + (x - t_i)[C_i + (x - t_i)[B_i + (x - t_i)A_i]] \tag{11}$$

其中

$$A_i = \frac{1}{6h_i}(z_{i+1} - z_i)$$

$$B_i = \frac{z_i}{2}$$

$$C_i = -\frac{h_i}{6}z_{i+1} - \frac{h_i}{3}z_i + \frac{1}{h_i}(y_{i+1} - y_i)$$

根据以上解说，读者不难编写出一个子程序或过程来执行对(11)式的计算．它输入整数n，结点数组(t_i)，函数值数组(y_i)以及由前面的程序得到的数组(z_i)，它还可以输入一个实数x，运行后给出$S(x)$的值．

下面是对上述算法的两个程序作简单的计算机测试后的某些结果．我们选取函数$f(x)=$

$\sqrt{x}$，它在区间[0，2.25]内的10个等距结点上用三次样条函数 S 插值，打印出37个结点上的误差值 $E(x)\equiv S(x)-f(x)$

x	$\mid E(x)\mid$
0.000 0	0.0
0.062 5	$1.073\,21\times10^{-1}$
0.125 0	$7.526\,66\times10^{-2}$
0.187 5	$3.326\,17\times10^{-2}$
0.250 0	0.0
⋮	⋮
1.750 0	0.0
1.812 5	$3.647\,80\times10^{-5}$
1.875 0	$6.365\,78\times10^{-5}$
1.937 5	$5.853\,18\times10^{-5}$
2.000 0	0.0
2.062 5	$1.140\,83\times10^{-4}$
2.125 0	$2.123\,12\times10^{-4}$
2.187 5	$2.046\,82\times10^{-4}$
2.250 0	0.0

输出信息表明在每个结点上的误差 E 是零. 在异于结点的其他点上，$\mid E\mid$ 不超过0.11. $\mid E\mid$ 的最大值出现在区间$[t_0, t_1]$内.（为什么?）

现在给出一个定理，说明自然三次样条可能是最光滑的插值函数. 名词光滑在定义中被赋予了专门的意义.

354

定理1（自然三次样条最优性定理） 设 f'' 在$[a, b]$内连续并且 $a=t_0<t_1<\cdots<t_n=b$. 若 S 是 f 在结点 t_i 上的自然三次样条插值，$0\leqslant i\leqslant n$，则

$$\int_a^b[S''(x)]^2\mathrm{d}x\leqslant\int_a^b[f''(x)]^2\mathrm{d}x$$

证明 令 $g\equiv f-S$. 从而对于 $0\leqslant i\leqslant n$，$g(t_i)=0$ 并且

$$\int_a^b(f'')^2\mathrm{d}x=\int_a^b(S'')^2\mathrm{d}x+\int_a^b(g'')^2\mathrm{d}x+2\int_a^bS''g''\mathrm{d}x$$

如果我们能证明

$$\int_a^bS''g''\mathrm{d}x\geqslant0$$

则定理证毕. 我们在下面的分析中证明此式，利用分部积分法，条件 $S''(t_0)=S''(t_n)=0$，以及在$[t_{i-1}, t_i]$上 S''' 是常数（记为 c_i），我们有

$$\begin{aligned}\int_a^bS''g''\mathrm{d}x&=\sum_{i=1}^n\int_{t_{i-1}}^{t_i}S''g''\mathrm{d}x\\&=\sum_{i=1}^n\left\{(S''g')(t_i)-(S''g')(t_{i-1})-\int_{t_{i-1}}^{t_i}S'''g'\mathrm{d}x\right\}\end{aligned}$$

$$= -\sum_{i=1}^{n}\int_{t_{i-1}}^{t_i} S'''g'\,dx = -\sum_{i=1}^{n} c_i\int_{t_{i-1}}^{t_i} g'\,dx$$

$$= -\sum_{i=1}^{n} c_i[g(t_i) - g(t_{i-1})] = 0$$ ■

记得由 $y=f(x)$ 给出的曲线的曲率是

$$|f''(x)|[1+\{f'(x)\}^2]^{-3/2}$$

如果去掉括号中的非线性项，$|f''(x)|$ 是曲率的一个近似值．在自然三次样条插值中，因为 $\int_a^b [f''(x)]^2 dx$ 被最小化，所以我们找到一条在一个区间上具有最小(近似的)曲率的曲线．

在上述证明的步骤中，我们注意到有一个"坍缩"和：

355

$$\sum_{i=1}^{n}[(S''g')(t_i) - (S''g')(t_{i-1})] = (S''g')(b) - (S''g')(a)$$

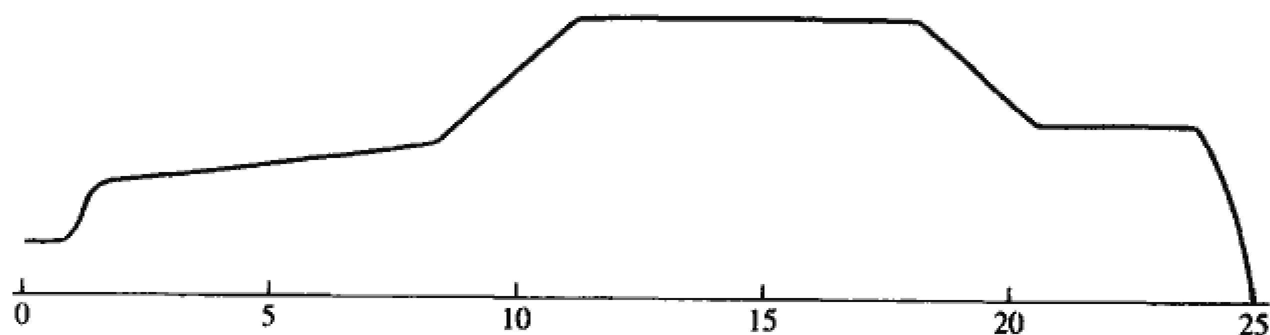
当最后的这个表达式非负时，我们的证明仍然是正确的．令 $g=f-S$，我们得到条件

$$S''(b)[f'(b) - S'(b)] \geqslant S''(a)[f'(a) - S'(a)]$$

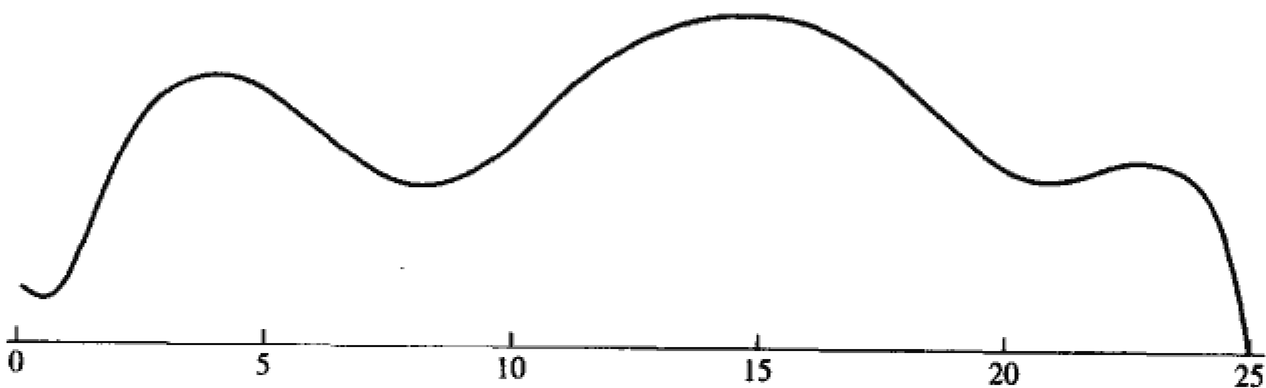
例如，如果我们假设端点条件 $S'(a)=f'(a)$ 和 $S'(b)=f'(b)$，用来替代假设 S 是自然三次样条插值，则结论也成立．

6.4.2 张力样条

在一些数据拟合问题中，得到一个称为**张力**的参数 τ 是很有用的，当 τ 取较大数值时，通过数据点的曲线将有较高的张力．这个 τ 可以解释为拉直数据点之间曲线的力，如图 6-5a 和 b 所示．



a) 高张力样条曲线($\tau=10$)



b) 低张力样条曲线($\tau=0.1$)

图 6-5

当 τ 取较小数值时，曲线将更接近三次样条插值的形状，当 $\tau \to +\infty$ 时，曲线近似于分段线性函数，即一次样条函数.

下面讨论上述曲线的一个数学模型. 与前面一样，我们有结点

$$t_0 < t_1 < \cdots < t_n$$

在每一点 t_i 上给定数据 y_i. 我们要寻找的**张力样条**就是具有下列性质的函数 f：

1. 函数 $f \in C^2[t_0, t_n]$.
2. 对于 $0 \leqslant i \leqslant n$，插值条件 $f(t_i) = y_i$ 成立.
3. 在每一个开区间 (t_{i-1}, t_i) 上，f 满足 $f^{(4)} - \tau^2 f'' = 0$.

356

因此，f 整体具有二次连续导数，插值给定的数据，并且在每一个子区间内都满足某一个微分方程. 因为方程 $f^{(4)} = 0$ 的解是三次多项式，所以当 $\tau = 0$ 时，上面的约定显然产生一个三次样条.

为确定 f，我们仿照三次样条的情况进行讨论. 因此，我们取 $z_i \equiv f''(t_i)$ 并且写出在区间 $[t_i, t_{i+1}]$ 上 f 必须满足的条件：

$$f^{(4)} - \tau^2 f'' = 0$$
$$f(t_i) = y_i \qquad f(t_{i+1}) = y_{i+1}$$
$$f''(t_i) = z_i \qquad f''(t_{i+1}) = z_{i+1}$$

可以证明这个**两点边值问题**的解是

$$f(x) = \{z_i \sinh[\tau(t_{i+1} - x)] + z_{i+1} \sinh[\tau(x - t_i)]\}/[\tau^2 \sinh(\tau h_i)] + (y_i - z_i/\tau^2)(t_{i+1} - x)/h_i + (y_{i+1} - z_{i+1}/\tau^2)(x - t_i)/h_i \tag{12}$$

在确定系数 z_i 的值以后，上述等式将用来计算 f 在区间 $[t_i, t_{i+1}]$ 上的值. 所有这一切都类似于三次样条的情况.

由于 f 具有 C^2 整体光滑的性质，所以条件

$$\lim_{x \downarrow x_i} f'(x) = \lim_{x \uparrow x_i} f'(x) \qquad (1 \leqslant i \leqslant n-1)$$

在内结点上必须成立. 这里我们就不给出其中所涉及的冗长乏味的计算过程了，它们可以仿照三次样条的情况那样处理，其结果是含未知量 $z_0, z_1, \cdots, z_n$ 的一个三对角方程组，它可写成下面的形式

$$\alpha_{i-1} z_{i-1} + (\beta_{i-1} + \beta_i) z_i + \alpha_i z_{i+1} = \gamma_i - \gamma_{i-1} \qquad (1 \leqslant i \leqslant n-1) \tag{13}$$

其中 α_i，β_i 和 γ_i 分别为：

$$\alpha_i = 1/h_i - \tau/\sinh(\tau h_i)$$
$$\beta_i = \tau \cosh(\tau h_i)/\sinh(\tau h_i) - 1/h_i$$
$$\gamma_i = \tau^2 (y_{i+1} - y_i)/h_i$$

可以看出，为了确定 z 向量，我们还需要两个附加条件. 和三次样条的情况一样，一种可能是指定 $z_0 = z_n = 0$.

确定一个拟合数据 (t_i, y_i) 的张力样条 f 可以分为以下几步：

1. 核实 $t_0 < t_1 < \cdots < t_n$.
2. 对于 $0 \leqslant i \leqslant n-1$，计算 h_i，α_i，β_i，γ_i.

3. 令 $z_1=z_n=0$.

4. 求解三对角方程组(13)中的 z_i，$2\leqslant i\leqslant n-1$.

[357] 5. 用(12)式计算 f 在区间$[t_i, t_{i+1}]$上的值.

我们所勾勒出的这个清晰易懂的方法可用于一个有趣的试验. 例如，图 6-5(a)和(b)所示的两条曲线，在第一条曲线中令 $\tau=10$，在第二条曲线中令 $\tau=0.1$. 对于 $0\leqslant i\leqslant 25$，在整数点 $t_i=i$ 上给定的函数值是相同的.

张力样条是 Schweikert[1966]引进的. Cline[1974a, b]和 Pruess[1976, 1978]也给出了一些相关的论文. Cline 研制出了利用张力样条计算曲线和曲面的软件. 另一类有关张力的样条是 de Boor[1984]提出的**套紧样条**. 这些函数是常规的三次样条，它在希望曲线突然改变方向的区域内增加结点(和数据). 套紧样条的好处是不需要新的计算机程序，并且可以避免使用双曲函数所带来的计算工作.

6.4.3 高次自然样条的理论

在本节的最后部分，我们介绍一些高次自然样条的理论. 因为只存在奇数次的自然样条，所以我们用 $2m+1$ 表示样条的次数. 当 $m=1$ 时，如前所述，我们有自然三次样条. 为方便起见，我们将用一种稍许不同的方式来介绍一般理论. 和前面一样，给定结点集如下：

$$t_0<t_1<\cdots<t_n$$

一个 $2m+1$ 次的自然样条是一个函数 $S\in C^{2m}(\mathbb{R})$，在每一个区间$[t_0, t_1]$, $[t_1, t_2]$, $\cdots$, $[t_{n-1}, t_n]$内，它都化为一个次数$\leqslant 2m+1$ 的多项式，而在区间$(-\infty, t_0)$和(t_n, ∞)内化为一个次数至多为 m 的多项式. 在此定义下，一个自然三次样条在区间$(-\infty, t_0)$和(t_n, ∞)内一定化为线性多项式.

我们将在 $n+1$ 个结点 $t_0, t_1, \cdots, t_n$ 上的全体 $2m+1$ 次自然样条所构成的线性空间记为 $\mathcal{N}^{2m+1}(t_0, t_1, \cdots, t_n)$，或简记为 $\mathcal{N}_n^{2m+1}$.

使用所谓的**截断幂函数**会很方便. 该函数记为 x_+^n，如果 $x\geqslant 0$ 它定义为 x^n，如果 $x<0$，它定义为 0. 它属于连续函数类 C^{n-1}.

定理 2(截断幂函数定理) $\mathcal{N}_n^{2m+1}$ 中的每个元有表达式

$$S(x)=\sum_{i=0}^{m}a_ix^i+\sum_{i=0}^{n}b_i(x-t_i)_+^{2m+1} \tag{14}$$

其中对于 $0\leqslant j\leqslant m$, $\sum_{i=0}^{n}b_it_i^j=0$.

证明 在区间$(-\infty, t_0)$内，S 是一个次数至多是 m 次的多项式 p_0. 这个多项式可确定系数 a_i. 在区间(t_0, t_1)内，S 化为一个 $2m+1$ 次多项式 p_1. 在点 t_0 的连续性条件是

[358]

$$p_0^{(j)}(t_0)=p_1^{(j)}(t_0)\qquad(0\leqslant j\leqslant 2m)$$

根据泰勒定理，p_1 可写成

$$p_1(x)=\sum_{j=0}^{2m+1}\frac{1}{j!}p_1^{(j)}(t_0)(x-t_0)^j$$

$$= \sum_{j=0}^{2m} \frac{1}{j!} p_0^{(j)}(t_0)(x-t_0)^j + b_0(x-t_0)^{2m+1}$$

$$= p_0(x) + b_0(x-t_0)^{2m+1}$$

这个等式表明，在$(-\infty, t_1)$上我们有

$$S(x) = p_0(x) + b_0(x-t_0)_+^{2m+1}$$

(对于$m<j\leqslant 2m$，我们会注意到$S^{(j)}(t_0)=0$)在每个结点上都重复进行点t_0上的讨论过程，就可以得到其余的项$b_i(x-t_i)_+^{2m+1}$. 在区间(t_n, ∞)上，S一定会化为一个次数$\leqslant m$的多项式. 因此，在这个区间内，

$$0 = S^{(m+1)}(x) = \sum_{i=0}^{n} b_i(2m+1)(2m)\cdots(m+1)(x-t_i)^m$$

根据这个等式以及借助于二项式定理，我们得到

$$0 = \sum_{i=0}^{n} b_i(x-t_i)^m = \sum_{i=0}^{n} b_i \sum_{j=0}^{m} \binom{m}{j} x^{m-j}(-t_i)^j$$

其中有一个次数至多是m次的x的多项式，它的系数一定是零. 因此，对于$j=0, 1, \cdots, m$，有$\sum_{i=0}^{n} b_i t_i^j = 0$. ■

定理3(奇数次自然样条唯一性定理) *给定结点$t_0<t_1<\cdots<t_n$，设$0\leqslant m\leqslant n$. 则存在唯一的$2m+1$次自然样条在这些结点上取到给定的值.*

证明 根据定理2，自然样条有下列形式

$$S(x) = \sum_{j=0}^{m} a_j x^j + \sum_{j=0}^{n} b_j(x-t_j)_+^{2m+1}$$

如果S的给定值是λ_i，那么(与定理2一样)插值问题就需要我们求解线性方程组

$$\begin{cases} S(t_i) \equiv \sum_{j=0}^{m} b_j t_j^i + \sum_{j=0}^{n} b_j(t_i-t_j)_+^{2m+1} = \lambda_i & (0\leqslant i\leqslant n) \\ \sum_{j=0}^{m} b_j t_j^i = 0 & (0\leqslant i\leqslant m) \end{cases}$$

这是一个含$m+n+2$个未知量$m+n+2$个方程的方程组. 为证明它是非奇异的，只要证明对应的齐次问题仅有零解即可. 因此，对于$0\leqslant i\leqslant n$，假设$S(t_i)=0$. 我们将证明 [359]

$$I \equiv \int_a^b [S^{(m+1)}(x)]^2 \mathrm{d}x = 0 \tag{15}$$

其中$a=t_0$及$b=t_n$. 由分部积分得到

$$I = S^{(m+1)}(x)S^{(m)}(x)\Big|_a^b - \int_a^b S^{(m)}(x)S^{(m+2)}(x)\mathrm{d}x$$

$$= -\int_a^b S^{(m)}(x)S^{(m+2)}(x)\mathrm{d}x$$

此处我们要用到事实：在$(-\infty, a)$内S是一个次数至多是m次的多项式，因此$S^{(m+1)}(a)=0$. 同样地，$S^{(m+1)}(b)=0$. 重复上述讨论直到我们得到

$$I=(-1)^m\int_a^b S^{(1)}(x)S^{(2m+1)}(x)\mathrm{d}x$$

因为 $S^{(2m+1)}$ 是分段常值函数，所以

$$I=(-1)^m\sum_{i=1}^n\int_{t_{i-1}}^{t_i}c_iS'(x)\mathrm{d}x=(-1)^m\sum_{i=1}^n c_i[S(t_i)-S(t_{i-1})]=0$$

这样就证明了(15)式成立．由此我们推断 $S^{(m+1)}\equiv0$．因此，S 是一个次数至多是 m 次的多项式．由于 S 有零点 t_0，t_1，…，t_n 并且 $n+1>m$，我们可知 $S=0$． ■

下面是自然三次样条光滑性定理的类似结果．

定理 4(*奇数次自然样条最优性定理*) *设 $m\leqslant n$ 及 $f\in C^{m+1}[a,b]$．假设 S 是在结点 $a=t_0<t_1<\cdots<t_n=b$ 上插值 f 的 $2m+1$ 次自然样条，则*

$$\int_a^b[S^{(m+1)}(x)]^2\mathrm{d}x\leqslant\int_a^b[f^{(m+1)}(x)]^2\mathrm{d}x$$

证明 类似于定理 1 的证明过程，令 $g=f-S$．那么对于 $0\leqslant i\leqslant n$，有 $g(t_i)=0$．重复分部积分法，可以证明

$$\int_a^b g^{(m+1)}(x)S^{(m+1)}(x)\mathrm{d}x=0$$

[360] 再利用下列过程即可完成该定理的证明

$$\begin{aligned}\int_a^b[f^{(m+1)}(x)]^2\mathrm{d}x&=\int_a^b[S^{(m+1)}(x)+g^{(m+1)}(x)]^2\mathrm{d}x\\&=\int_a^b[S^{(m+1)}(x)]^2\mathrm{d}x+2\int_a^b S^{(m+1)}(x)g^{(m+1)}(x)\mathrm{d}x+\int_a^b[g^{(m+1)}(x)]^2\mathrm{d}x\\&=\int_a^b[S^{(m+1)}(x)]^2\mathrm{d}x+\int_a^b[g^{(m+1)}(x)]^2\mathrm{d}x\\&\geqslant\int_a^b[S^{(m+1)}(x)]^2\mathrm{d}x\end{aligned}$$

■

习题 6.4

1. 参考求 z_i 值的三对角算法，证明对所有 $i=n-1$，$n-2$，…，1，有 $u_iz_i+h_iz_{i+1}-v_i=0$．
2. (续)用 E_i 表示上题中等式的左端项．因而$(h_i/u_i)E_i+E_{i+1}=0$．根据算法中的公式，证明后面的等式可以化为(10)式．这就确定算法产生了(10)式的一个解．
3. 证明：在(6)式中，我们若用 t_i+h_i 替换 t_{i+1}，则其结果是

$$S_i(x)=\frac{z_{i+1}}{6h_i}(x-t_i)^3-\frac{z_i}{6h_i}(x-t_i-h_i)^3+C(x-t_i)-D(x-t_i-h_i)$$

4. (续)展开上述等式中的项$(x-t_i-h_i)^3$ 并且利用 C 和 D 的正确值，证明课本中等式(11)是正确的．
5. 确定下列函数是否为一个二次样条函数：

$$f(x)=\begin{cases}x & x\in(-\infty,1]\\-\frac{1}{2}(2-x)^2+\frac{3}{2} & x\in[1,2]\\\frac{3}{2} & x\in[2,\infty)\end{cases}$$

6. (续)试问上题中的函数是一个三次样条函数吗？

7. 确定 a, b, c, d, e 的值，使得下列函数是一个三次样条：

$$f(x)=\begin{cases}a(x-2)^2+b(x-1)^3 & x\in(-\infty,1]\\ c(x-2)^2 & x\in[1,3]\\ d(x-2)^2+e(x-3)^3 & x\in[3,\infty)\end{cases}$$

其次，确定这些参数的值，使得这个三次样条插值下列表值：

x	0	1	4
y	26	7	25

8. 证明课本中的(7)式，(9)式和(10)式.

9. 以三次样条的讨论过程为蓝本，推导适当的公式和算法，给出一个二次样条插值数据$(t_i,\ y_i)$，$0\leqslant i\leqslant n$，其中 $t_0<t_1<\cdots<t_n$. 如果 Q 是这个样条插值，那么数值 $z_i=Q'(t_i)$有意义. 求出适应于 z_0，z_1，…，z_n 的方程. 你将会发现其中的一个 z 点可以是任意的，例如 $z_0=0$.

10. 证明：在求解方程组(10)的算法中，对所有 $i=1$，2，…，$n-1$，都有 $u_i>h_i+h_{i-1}$. 361

11. 确定 a, b, c 的值，使得下列函数是一个具有结点 0，1，2 的三次样条：

$$f(x)=\begin{cases}3+x-9x^2 & x\in[0,1]\\ a+b(x-1)+c(x-1)^2+d(x-1)^3 & x\in[1,2]\end{cases}$$

其次，确定 d 使得 $\int_0^2[f''(x)]^2\mathrm{d}x$ 达到极小. 最后，求 d 的一个值，使得 $f''(2)=0$，并解释为什么这个值不同于前面所确定的值.

12. 确定下列函数是否为一个三次样条

$$f(x)=\begin{cases}x^3+x & x\leqslant 0\\ x^3-x & x\geqslant 0\end{cases}$$

证明

$$\lim_{x\uparrow 0}f''(x)=\lim_{x\downarrow 0}f''(x)$$

13. 确定插值表值

x	0	1	2	3
y	1	1	0	10

的自然三次样条是否为下列函数.

$$f(x)=\begin{cases}1+x-x^3 & x\in[0,1]\\ 1-2(x-1)-3(x-1)^2+4(x-1)^3 & x\in[1,2]\\ 4(x-2)+9(x-2)^2-3(x-2)^3 & x\in[2,3]\end{cases}$$

14. 确定下列函数是否为一个自然三次样条.

$$f(x)=\begin{cases}2(x+1)+(x+1)^3 & x\in[-1,0]\\ 3+5x+3x^2 & x\in[0,1]\\ 11+11(x-1)+3(x-1)^2-(x-1)^3 & x\in[1,2]\end{cases}$$

15. 微积分中有定理断言：若一个函数在一点可微，则它必在该点连续. 其简单原因如下：若极限定义 $f'(x)$ 存在，即

$$f'(x)=\lim_{h\to 0}\frac{f(x+h)-f(x)}{h}$$

则分式中分子的极限一定是 0，由此推得 f 在点 x 处的连续性. 现在假设对某一函数 f 和某一点 x_0，有

$$\lim_{x\downarrow x_0}f'(x)=\lim_{x\uparrow x_0}f'(x)$$

那么我们可以推断 f' 在点 x_0 处连续吗？

16.(续)在检验一个分段三次样条函数 f 是否为一个三次样条时，对每个结点 t 证明等式

$$\lim_{x\downarrow t} f''(x) = \lim_{x\uparrow t} f''(t)$$

成立就足够了吗？

362 17. 试求一个自然三次样条函数 S，它的结点是 -1，0，1，并且取值分别是 $S(-1)=13$，$S(0)=7$，$S(1)=9$.

18. 试问函数

$$f(x) = \begin{cases} (x+1)+(x+1)^3 & x \in [-1,0] \\ 4+(x-1)+(x-1)^3 & x \in [0,1] \end{cases}$$

具有自然三次样条的哪些性质？它不具有自然三次样条的哪些性质？

19. 试求出一个自然三次样条函数，它的结点是 -1，0，1，并且取值是

x	-1	0	1
y	5	7	9

20. 确定是否存在系数 a，b，c，d，使得函数

$$S(x) = \begin{cases} 1-2x & x \in (-\infty,-3] \\ a+bx+cx^2+dx^3 & x \in [-3,4] \\ 157-32x & x \in [4,+\infty) \end{cases}$$

是区间 $[-3, 4]$ 上的一个自然三次样条.

21. 下列函数是一个自然三次样条吗？

$$S(x) = \begin{cases} x^3-1 & x \in \left[-1,\frac{1}{2}\right] \\ 3x^3-1 & x \in \left[\frac{1}{2},1\right] \end{cases}$$

22.(续)对下列函数重复上题中的问题.

$$S(x) = \begin{cases} x^3-1 & x \in [-1,0] \\ 3x^3-1 & x \in [0,1] \end{cases}$$

23. 能否确定 a 和 b 使得函数

$$S(x) = \begin{cases} (x-2)^3+a(x-1)^2 & x \in (-\infty,2] \\ (x-2)^3-(x-3)^2 & x \in [2,3] \\ (x-3)^3+b(x-2)^2 & x \in [3,+\infty) \end{cases}$$

是一个自然三次样条？试问为什么能或者不能？

24. 如果 S 是在结点序列 $0=t_0<t_1<\cdots<t_n=1$ 上插值 f 的一次样条函数，试问 $\int_0^1 S(x)\mathrm{d}x$ 是什么？

25. 试问 (a, b, c, d) 取什么值可使得下列函数是一个三次样条？

$$f(x) = \begin{cases} x^3 & x \in [-1,0] \\ a+bx+cx^2+dx^3 & x \in [0,1] \end{cases}$$

26. 确定 (a, b, c) 的值，使得函数

$$f(x) = \begin{cases} x^3 & x \in [0,1] \\ \frac{1}{2}(x-1)^3+a(x-1)^2+b(x-1)+c & x \in [1,3] \end{cases}$$

是一个三次样条. 请问它是一个自然三次样条吗？

27. 设 $t_0<t_1<\cdots<t_n$ 并且 $-\infty<x<+\infty$. 试问下列算法输出的 k 值是什么？

```
for i=1 to n do
  if x<t_i then
      k←i
    exit loop
  end if
end do
```

363

28. 在点 $a=t_0<t_1<\cdots<t_{20}=b$ 上定义一个二次样条插值函数 $S(x)$ 需要多少个条件？$S'(x)$ 的连续性能提供所需要的全部条件吗？
29. 当 S 的端点条件变成 $S'(a)=f'(a)$ 和 $S'(b)=f'(b)$ 时，证明定理 1.
30. 当端点条件具体指定为 $S'(t_0)$ 和 $S'(t_n)$ 的值时，给出一个适当的程序用于寻求三次样条插值.
31. 给出适合等距结点情况的自然三次样条插值的简化形式.
32. 证明：具有结点 $t_0<t_1<\cdots<t_n$ 的一次样条函数可表示为下列形式

$$S(x) = ax + b + \sum_{i=1}^{n-1} c_i \mid x - t_i \mid$$

33. 证明：(12)式中的函数是(12)式前面所给出的两点边值问题的解.
34. 证明：如课本中所述，由整体光滑条件可知(13)式成立.
35. 证明：(13)式中的系数矩阵是对角占优的.
36. 用定理 4 的记号和假设条件，证明不等式

$$\| f^{(m+1)} \| \geqslant \| f^{(m+1)} - S^{(m+1)} \|$$

成立.

计算机习题 6.4

1. 在一张绘图纸上画一条曲线，例如一条卵形线或一条螺线. 沿曲线选择一些较规则分布的点，给它们标号为 $t_0=1.0$，$t_1=2.0$，等等. 写出每一个选择点的 x 坐标和 y 坐标，得到 $x(t)$ 和 $y(t)$ 的表值. 用样条函数 S 和 S^* 拟合这些函数. 因而公式 $x=S(t)$ 和 $y=S^*(t)$ 给出曲线的一个近似参数表达式. 用自动绘图仪绘制出几种不同测试情况的曲线.
2. 在习题 6.4.9 的讨论中，找出一种方法，求解数组 $(z_0, z_1, \cdots, z_n)$ 使得 $\sum_{i=0}^{n} z_i^2$ 达到极小. 也就是说，用这个条件替换那个习题中的任意性条件. 把这个特征纳入你的算法并在计算机上进行测试.
3. 证明公式

$$\int_{t_i}^{t_{i+1}} S_i(x)\mathrm{d}x = \frac{h_i}{2}(y_i + y_{i+1}) - \frac{h_i^3}{24}(z_i + z_{i+1})$$

然后编写并测试一个程序用于计算

$$\int_{t_0}^{t_n} S(x)\mathrm{d}x$$

364

4. 编写并测试一个计算机程序，用于计算具有给定结点 $t_0<t_1<\cdots<t_n$ 并且满足下列条件的三次样条函数 S：

$$\begin{cases} S(t_i) = y_i \quad (0 \leqslant i \leqslant n) \\ S''(t_0) = \alpha \\ S''(t_n) = \beta \end{cases}$$

5. 为张力样条的算法编写程序，并用不同的张力参数 τ 的值测试这个程序.
6. 从图 6-6 中所示的轿车车体的轮廓开始，准备 10 到 20 各点横坐标和纵坐标的一个表. 用张力值 $\tau=0.25$，4，10 产生并绘制出张力样条插值的值. 观察哪一个结果产生的外观图形最令人满意.

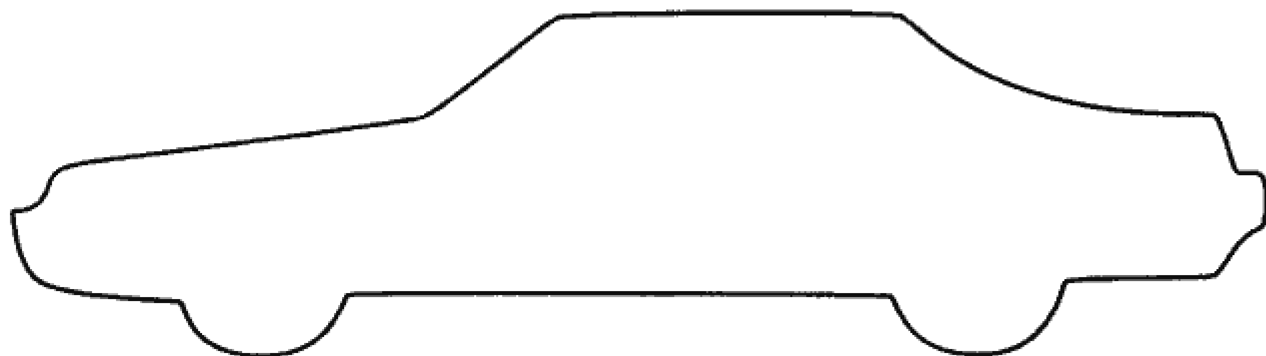


图 6-6 轿车轮廓

7. 如图 6-7 所示，画一个手写体字母．然后，借助于三次样条和绘图仪重新画一次．具体操作如下：在曲线上选择较合理数目的点，例如 $n=11$．给这些点标号 $t=1, 2, \cdots, n$．对每一点，求出对应的 x 坐标和 y 坐标．用三次样条插值函数 S_x 和 S_y，拟合 $x=S_x(t)$ 和 $y=S_y(t)$．这将会产生原曲线的一个参数表达式．计算出大量的 $S_x(t)$ 和 $S_y(t)$ 的值并且提供给绘图仪．要想更多地了解有关样条曲线如何应用于字体设计方面的内容，读者可以查阅 Knuth [1979]．

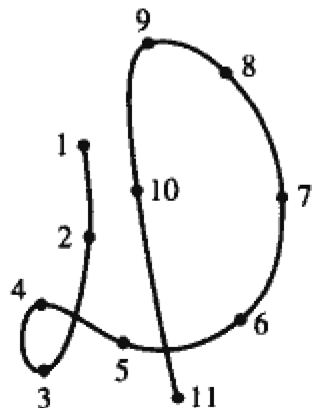


图 6-7 11 个结点的手写体字母

8. 解释下列数值试验的结果并得出一些相应的结论．
   a. 定义 p 是一个 20 次多项式，它在区间$[-1, 1]$内的 21 个等距结点上的插值函数 $f(x)=(1+6x^2)^{-1}$．区间端点包括在结点中．打印出 $f(x)$，$p(x)$，$f(x)-p(x)$在该区间 41 个等距结点上的表值．

   365

   b. 用下列切比雪夫结点重复上述试验．
   $$x_i = \cos[(i-1)\pi/20] \qquad (1 \leqslant i \leqslant 21)$$
   c. 在 21 个等距结点上，利用一个三次插值样条重复上述试验．

6.5 *B* 样条：基本理论

这一节专门讨论样条函数系统，使得其他所有样条函数都可以由它的线性组合得到．这些样条构成某些样条空间的基．所以称为 ***B* 样条**．一旦给定结点，B 样条就很容易通过递归关系产生．而且算法也比较简单．B 样条以其优美的理论和数值计算中的典型性质著称．此外，B 样条还可以得到进一步的推广．

我们从实轴上一组结点序列 t_i 开始，为了实用的目的，无论何时也只需要有限个结点，但是如果把结点集合左端扩展到$-\infty$，右端扩展到$+\infty$，使其成为一个无限集：

$$\cdots < t_{-2} < t_{-1} < t_0 < t_1 < t_2 < \cdots$$

可以使得理论上讨论起来更加容易．在本节中，我们假设这个结点序列是固定的，并在此基础上建立所有的样条．

6.5.1 0 次 *B* 样条

0 次 B 样条记为 B_i^0，图形如图 6-8 所示．指标 i 取遍全体整数．图形中的深色圆点表明我们定义 $B_i^0(t_i)=1$ 以及 $B_i^0(t_{i+1})=0$．正式的定义是：

$$B_i^0(x) = \begin{cases} 1 & 若\ t_i \leqslant x < t_{i+1} \\ 0 & 其他 \end{cases}$$

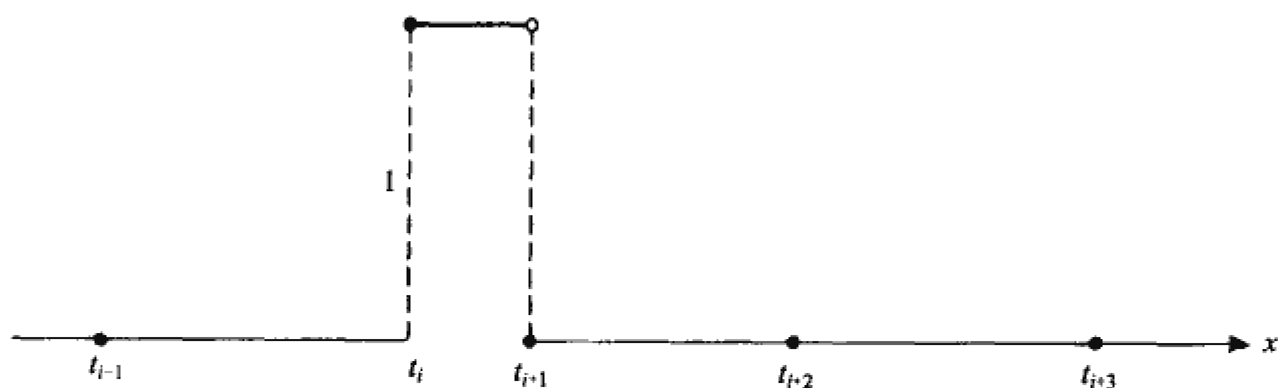


图 6-8 B 样条 B_i^0

366

这些 B 样条构成一个无穷序列，$\{B_i^0 : i\in \mathbb{Z}\}$.（$\mathbb{Z}$ 在这里表示全体整数的集合：正整数，负整数和 0.）可以看出它们的一些显而易见的性质：

1. 使得 $B_i^0(x)\neq 0$ 的 x 的集合定义为 B_i^0 的**支撑**，它是区间$[t_i, t_{i+1})$.
2. 对一切 i 及 x，$B_i^0(x)\geqslant 0$.
3. 在整个实轴上 $B_i^0(x)$是右连续的.
4. 对一切 x，$\sum\limits_{i=-\infty}^{\infty} B_i^0(x) = 1$.

我们证明上述最后一个等式. 任取 $x\in\mathbb{R}$，然后确定 x 所在的结点区间，例如，$t_j\leqslant x<t_{j+1}$；因而 $\sum\limits_{i=-\infty}^{\infty} B_i^0(x) = B_j^0(x) = 1$.

最后，关于样条 B_i^0 还要说明一点，对于给定结点序列上的全体 0 次样条，假设我们标准化这些样条使其都是右连续的，那么样条 B_i^0 构成全体 0 次样条的一个基. 为了证明这个断言，我们假设 S 是一个右连续的 0 次样条函数. 因而它是一个分段常值函数，并且由如下形式的一组规则来定义：

$$S(x) = c_i \quad 若 \quad t_i\leqslant x<t_{i+1} \qquad (i\in\mathbb{Z})$$

显然有 $S(x) = \sum\limits_{i=-\infty}^{\infty} c_i B_i^0(x)$.（因此，我们有 Schauder 意义下的一个基：空间中的每个向量有唯一一个无穷级数 $\sum\limits_{i=-\infty}^{\infty} c_i B_i^0$ 形式的表达式.）

函数 B_i^0 是所有高次 B 样条递归定义的出发点. 基本的递归关系是

$$B_i^k(x) = \left(\frac{x-t_i}{t_{i+k}-t_i}\right)B_i^{k-1}(x) + \left(\frac{t_{i+k+1}-x}{t_{i+k+1}-t_{i+1}}\right)B_{i+1}^{k-1}(x) \qquad (k\geqslant 1) \tag{1}$$

高阶 B 样条的所有性质都将来自这个递归定义. 通过引入某些特殊的线性函数：

$$V_i^k(x) = \frac{x-t_i}{t_{i+k}-t_i} \tag{2}$$

我们可以把递归关系写成下列更优美的形式：

$$B_i^k = V_i^k B_i^{k-1} + (1-V_{i+1}^k)B_{i+1}^{k-1} \tag{3}$$

因为 B_i^0 是一个 0 次分段多项式，并且因为 V_i^k 是线性的，所以 B_i^1 是一个次数$\leqslant 1$ 的分段多项式. 同理表明，一般情况下 B_i^k 是一个次数$\leqslant k$ 的分段多项式.

6.5.2 一次 B 样条

借助(1)式，可以给出 $B_i^1(x)$ 的显式公式如下：

$$B_i^1(x) = \left(\frac{x-t_i}{t_{i+1}-t_i}\right)B_i^0(x) + \left(\frac{t_{i+2}-x}{t_{i+2}-t_{i+1}}\right)B_{i+1}^0(x)$$

$$= \begin{cases} 0 & \text{若 } x < t_i \text{ or } x \geqslant t_{i+2} \\ \dfrac{x-t_i}{t_{i+1}-t_i} & \text{若 } t_i \leqslant x < t_{i+1} \\ \dfrac{t_{i+2}-x}{t_{i+2}-t_{i+1}} & \text{若 } t_{i+1} \leqslant x < t_{i+2} \end{cases}$$

367 $B_i^1(x)$ 的图形如图 6-9 所示.

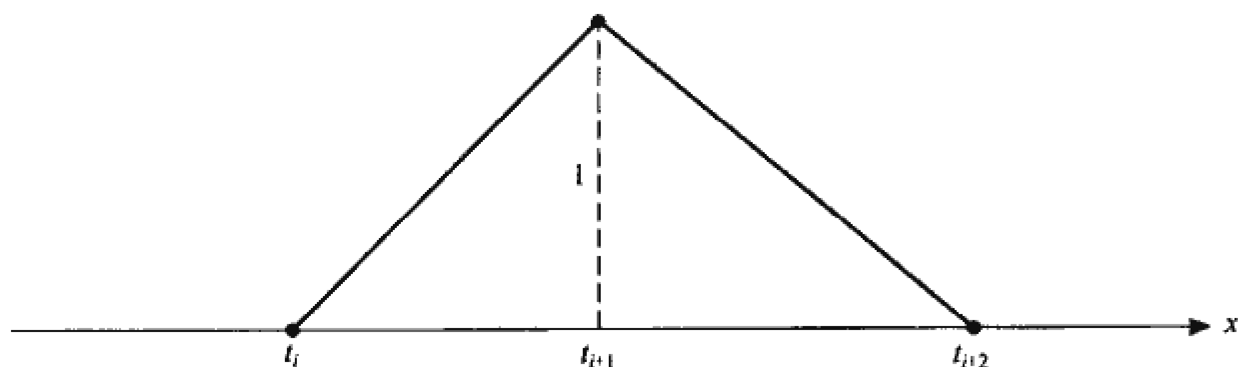


图 6-9　B 样条 B_i^1

此外还可以看出函数 B_i^1 的性质：

1. B_i^1 的支撑是 $(t_i,\ t_{i+2})$.
2. 对一切 i 及 x，$B_i^1(x) \geqslant 0$.
3. B_i^1 连续并且在除点 t_i，t_{i+1}，t_{i+2} 之外的每一点可微.
4. 对一切 x，$\sum_{i=-\infty}^{\infty} B_i^1(x) = 1$.

我们证明上述最后一个等式. 任取 $x \in \mathbb{R}$，因为当 i 增大时 t_j 收敛到 $+\infty$，当 i 减小时，t_j 收敛到 $-\infty$，所以我们可以找到一个指标 j，使得 $t_j \leqslant x < t_{j+1}$. 因而对于除 $i=j$ 或 $i=j-1$ 之外的所有 i，都有 $B_i^1(x)=0$. 因此，对这个 x，

$$\sum_{i=-\infty}^{\infty} B_i^1(x) = B_{j-1}^1(x) + B_j^1(x)$$

$$= \frac{t_{j+1}-x}{t_{j+1}-t_j} + \frac{x-t_j}{t_{j+1}-t_j} = 1$$

6.5.3 B 样条的性质

下面通过一系列引理，我们将给出函数族 $B_i^k(i \in \mathbb{Z},\ k \in \mathbb{N})$ 的重要性质.

引理 1(*B 样条的支撑引理*)　*若 $k \geqslant 1$ 并且 $x \notin (t_i,\ t_{i+k+1})$，则 $B_i^k(x)=0$.*

证明　我们已知对于 $k=1$ 结论成立，但对于 $k=0$ 结论不成立. 假设对于某个指标 $k-1$ 上述结论成立，那么对于 k 它也成立. 其理由如下：若 $x \notin (t_i,\ t_{i+k+1})$，则 $x \notin (t_i,\ t_{i+k})$ 并且 $x \notin$

(t_{i+1}, t_{i+k+1}). 根据归纳假设，$B_i^{k-1}(x)=0$ 并且 $B_{i+1}^{k-1}(x)=0$. 由(3)式知 $B_i^k(x)=0$ 成立. ■

引理 2(*B* 样条的正性引理) 设 $k\geqslant 0$. 若 $x\in(t_i, t_{i+k+1})$，则 $B_i^k(x)>0$.

证明 我们容易看出当 $k=0$ 或 $k=1$ 时，引理 2 结论成立.（根据前文中给出的 B_i^1 的显式公式知 $k=1$ 时结论成立.）假设对指标 $k-1$ 结论成立，$k\geqslant 2$. 对于一切 i 和 x，由归纳假设和引理 1 知 $B_i^k(x)\geqslant 0$. 设 $t_i<x<t_{i+k+1}$，因而(1)式右端项中的线性因式是正的. 再由归纳假设知，在 (t_i, t_{i+k}) 中 $B_i^{k-1}(x)>0$，在 (t_{i+1}, t_{i+k+1}) 中 $B_{i+1}^{k-1}(x)>0$. 因为 $k\geqslant 2$，所以这两个区间相互重叠，由(1)式就可以看出 $B_i^k(x)>0$. ■

368

因为我们希望用 *B* 样条 B_i^k 作为所有 k 次样条的基，所以我们会对形如 $\sum_{i=-\infty}^{\infty} c_i B_i^k(x)$ 的线性组合产生兴趣.

引理 3(*B* 样条的递归关系引理) 对一切 $k\geqslant 0$，我们有

$$\sum_{i=-\infty}^{\infty} c_i B_i^k = \sum_{i=-\infty}^{\infty} [c_i V_i^k + c_{i-1}(1-V_i^k)] B_i^{k-1}$$

证明 我们利用(3)式和基本的级数操作如下：

$$\sum_{i=-\infty}^{\infty} c_i B_i^k = \sum_{i=-\infty}^{\infty} c_i [V_i^k B_i^{k-1} + (1-V_{i+1}^k) B_{i+1}^{k-1}]$$

$$= \sum_{i=-\infty}^{\infty} c_i V_i^k B_i^{k-1} + \sum_{i=-\infty}^{\infty} c_{i-1}(1-V_i^k) B_i^{k-1}$$ ■

6.5.4 数值计算过程

在引理 3 中，系数 c_i 可以是常数或者函数. 因此，该引理提供了一个方法用来计算如下形式的函数：

$$f(x) = \sum_{i=-\infty}^{\infty} C_i^k(x) B_i^k(x)$$

我们假设函数 C_i^k 已经给定；当然它们可能是常数. 现在定义

$$C_i^{k-1}(x) = C_i^k(x) V_i^k(x) + C_{i-1}^k(x)[1-V_i^k(x)] \tag{4}$$

根据引理 3 和(4)式，我们有

$$\sum_{i=-\infty}^{\infty} C_i^k(x) B_i^k(x) = \sum_{i=-\infty}^{\infty} C_i^{k-1}(x) B_i^{k-1}(x)$$

对于 $k-1$，$k-2$，…，0，重复上述讨论，最终得到

369

$$\sum_{i=-\infty}^{\infty} C_i^k(x) B_i^k(x) = \sum_{i=-\infty}^{\infty} C_i^0(x) B_i^0(x)$$

正如我们所知，上式右端的表达式很容易计算：对于 $t_j\leqslant x<t_{j+1}$，它的值是 $C_j^0(x)$. 利用(2)式，可给出(4)式的详细表达式如下：

$$C_i^{j-1}(x) = [(x-t_i)C_i^j(x) + (t_{i+j}-x)C_{i-1}^j(x)]/(t_{i+j}-t_i) \tag{5}$$

上述评注导致出下列数值计算过程.

算法 1(*B* 样条系数算法) 如果给定系数 C_i^k，对于给定的 x，样条函数 $S(x)=\sum_{i=-\infty}^{\infty} C_i^k B_i^k(x)$ 可计算如下：确定指标 m 使得 $t_m\leqslant x<t_{m+1}$. 利用(5) 式，计算三角形数组

$$\begin{array}{ccccc} C_m^k & C_m^{k-1} & \cdots & C_m^1 & C_m^0 \\ C_{m-1}^k & C_{m-1}^{k-1} & \cdots & C_{m-1}^1 & \\ \vdots & \vdots & \iddots & & \\ C_{m-k+1}^k & C_{m-k+1}^{k-1} & & & \\ C_{m-k}^k & & & & \end{array}$$

因而 $S(x)=C_m^0$.

引理 4(*B* 样条单位分解引理) 对于一切 k，我们有

$$\sum_{i=-\infty}^{\infty} B_i^k(x) = 1$$

证明 利用刚才所述的算法来证明. 我们从 $\sum\limits_{i=-\infty}^{\infty} C_i^k B_i^k(x)$ 开始，其中对一切 i，$C_i^k = 1$，然后固定 x 并且用(4)式计算：

$$\begin{aligned} C_i^{k-1} &= C_i^k V_i^k + C_{i-1}^k[1-V_i^k] \\ &= V_i^k + 1 - V_i^k \\ &= 1 \end{aligned}$$

重复上述讨论，我们发现：对于一切 i 和 $j=k$，$k-1$，…，0，都有 $C_i^j=1$，因此，我们有

$$\sum_{i=-\infty}^{\infty} B_i^k(x) = \sum_{i=-\infty}^{\infty} B_i^0(x) = 1$$

■

6.5.5 *B* 样条的导数和积分

370 下一个引理给出 B_i^k 导数的一个重要公式. 为方便起见，我们用到(2)式中的 V_i^k 并且令

$$\alpha_i^k = \frac{1}{t_{i+k}-t_i} \tag{6}$$

利用这个记号，我们注意到

$$\frac{\mathrm{d}}{\mathrm{d}x} V_i^k(x) = \alpha_i^k \tag{7}$$

其他几个不证自明的有用公式是：

$$\alpha_i^k V_i^{k+1} = \alpha_i^{k+1} V_i^k \tag{8}$$

$$\alpha_{i+1}^k (1-V_i^{k+1}) = \alpha_i^{k+1}(1-V_{i+1}^k) \tag{9}$$

对上述两种情况，仅利用 α_i^k 和 V_i^k 的定义便可证明.

引理 5(*B* 样条导数引理) 对于 $k\geqslant 2$，*B* 样条函数的导数可如下计算：

$$\frac{\mathrm{d}}{\mathrm{d}x} B_i^k(x) = \left(\frac{k}{t_{i+k}-t_i}\right) B_i^{k-1}(x) - \left(\frac{k}{t_{i+k+1}-t_{i+1}}\right) B_{i+1}^{k-1}(x) \tag{10}$$

当 $k=1$ 时，该等式对于除 $x=t_i$，t_{i+1}，t_{i+2} 之外的一切 x 都成立.

证明 用数学归纳法证明. 把 $k=1$ 和 $k=2$ 时的情况留给读者完成. 我们假设对一固定的

k 公式成立. 在此假设之下，我们来证实下一个情况. 归纳假设的是(10)式，利用前面所说明的记号把它重写为紧凑形式：

$$\frac{\mathrm{d}}{\mathrm{d}x}B_i^k = k\alpha_i^k B_i^{k-1} - k\alpha_{i+1}^k B_{i+1}^{k-1} \tag{11}$$

根据基本递归关系(3)式，我们有

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}B_i^{k+1} &= \frac{\mathrm{d}}{\mathrm{d}x}[V_i^{k+1}B_i^k + (1-V_{i+1}^{k+1})B_{i+1}^k] \\ &= V_i^{k+1}\frac{\mathrm{d}}{\mathrm{d}x}B_i^k + \alpha_i^{k+1}B_i^k + (1-V_{i+1}^{k+1})\frac{\mathrm{d}}{\mathrm{d}x}B_{i+1}^k - \alpha_{i+1}^{k+1}B_{i+1}^k \end{aligned}\tag{12}$$

其中两次用到了(7)式. 接下来，我们归纳假设(11)式替换出现在(12)式右端的导数. 其结果是

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}B_i^{k+1} =& V_i^{k+1}(k\alpha_i^k B_i^{k-1} - k\alpha_{i+1}^k B_{i+1}^{k-1}) + \alpha_i^{k+1}B_i^k \\ &+ (1-V_{i+1}^{k+1})(k\alpha_{i+1}^k B_{i+1}^{k-1} - k\alpha_{i+2}^k B_{i+2}^{k-1}) - \alpha_{i+1}^{k+1}B_{i+1}^k\end{aligned}$$

通过简单的重新组合，上式可以写成下列形式

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}B_i^{k+1} =& \alpha_i^{k+1}B_i^k + k\alpha_i^k V_i^{k+1}B_i^{k-1} - \alpha_{i+1}^{k+1}B_{i+1}^k \\ &- k\alpha_{i+2}^k(1-V_{i+1}^{k+1})B_{i+2}^{k-1} - k\alpha_{i+1}^k V_i^{k+1}B_{i+1}^{k-1} \\ &+ k\alpha_{i+1}^k(1-V_{i+1}^{k+1})B_{i+1}^{k-1}\end{aligned}\tag{13}$$

371

我们将对这个等式做下列替换，那么所有的证明都可以利用(8)式和(9)式来完成：

$$\begin{aligned} k\alpha_i^k V_i^{k+1}B_i^{k-1} &= k\alpha_i^{k+1}V_i^k B_i^{k-1} \\ -k\alpha_{i+2}^k(1-V_{i+1}^{k+1})B_{i+2}^{k-1} &= -k\alpha_{i+1}^{k+1}(1-V_{i+2}^k)B_{i+2}^{k-1} \\ &\quad - k\alpha_{i+1}^k V_i^{k+1}B_{i+1}^{k-1} + k\alpha_{i+1}^k(1-V_{i+1}^{k+1})B_{i+1}^{k-1} \\ &= k\alpha_{i+1}^k(1-V_i^{k+1})B_{i+1}^{k-1} - k\alpha_{i+1}^k V_{i+1}^{k+1}B_{i+1}^{k-1} \\ &= k\alpha_i^{k+1}(1-V_{i+1}^k)B_{i+1}^{k-1} - k\alpha_{i+1}^{k+1}V_{i+1}^k B_{i+1}^{k-1}\end{aligned}$$

现在(13)式变换后的形式为：

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}B_i^{k+1} =& \alpha_i^{k+1}B_i^k + k[\alpha_i^{k+1}V_i^k B_i^{k-1} + \alpha_i^{k+1}(1-V_{i+1}^k)B_{i+1}^{k-1}] \\ &- \alpha_{i+1}^{k+1}B_{i+1}^k - k[\alpha_{i+1}^{k+1}V_{i+1}^k B_{i+1}^{k-1} + \alpha_{i+1}^{k+1}(1-V_{i+2}^k)B_{i+2}^{k-1}]\end{aligned}\tag{14}$$

利用基本的递归关系(3)式，化简(14)式带括号的项，得到

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}B_i^{k+1} &= \alpha_i^{k+1}B_i^k + k\alpha_i^{k+1}B_i^k - \alpha_{i+1}^{k+1}B_{i+1}^k - k\alpha_{i+1}^{k+1}B_{i+1}^k \\ &= (k+1)\alpha_i^{k+1}B_i^k - (k+1)\alpha_{i+1}^{k+1}B_{i+1}^k\end{aligned}\tag{15}$$

因为这个等式是用 $k+1$ 替换了(11)式中的 k，所以归纳证明步骤完成. ■

引理 6(*B 样条光滑性引理*) *对于 $k\geqslant 1$，B 样条 B_i^k 属于连续函数类 $C^{k-1}(\mathbb{R})$.*

证明 B_i^1 连续是显然的：$B_i^1\in C^0(\mathbb{R})$. 现在我们假设 $B_i^k\in C^{k-1}(\mathbb{R})$. 由引理 5 知，$(\mathrm{d}/\mathrm{d}x)B_i^{k+1}\in C^{k-1}(\mathbb{R})$. 因为这个导数是 B_i^k 和 B_{i+1}^k 的线性组合，所以 $B_i^{k+1}\in C^k(\mathbb{R})$. 根据归纳法原理，定理得证. ■

从与 B 样条导数相关的引理，我们得到一个有用的公式

$$\frac{\mathrm{d}}{\mathrm{d}x}\sum_{i=-\infty}^{\infty} c_i B_i^k(x) = k\sum_{i=-\infty}^{\infty}\left(\frac{c_i - c_{i-1}}{t_{i+k} - t_i}\right)B_i^{k-1}(x) \qquad (k \geqslant 2) \tag{16}$$

这个公式可以用于数值微分，尽管噪声信息的数值微分是一项非常不确定的任务. 但是当 $k=1$ 时，对于除结点之外的所有 x，(16)式都成立.

372

引理 7(*B* 样条积分引理) *B* 样条函数的积分可如下计算：

$$\int_{-\infty}^{x} B_i^k(s)\mathrm{d}s = \left(\frac{t_{i+k+1} - t_i}{k+1}\right)\sum_{j=i}^{\infty} B_j^{k+1}(x)$$

证明 根据(16)式以及

$$c_j = \begin{cases} 0 & 若\ j < i \\ 1 & 若\ j \geqslant i \end{cases}$$

我们可以证明引理中等式两端的导数相等. 因而除 $j=i$ 之外 $c_j - c_{j-1}$ 都为 0，并且我们得到

$$B_i^k(x) = \left(\frac{t_{i+k+1} - t_i}{k+1}\right)(k+1)\left(\frac{1}{t_{i+k+1} - t_i}\right)B_i^k(x)$$

可以确定等式两端的函数只相差一个常数，并且在点 $x=t_i$ 处两者都变为 0. ■

6.5.6 附加性质

如果 f 是一个函数并且 K 是它定义域内的一个子集，则 $f \mid K$ 表示 f 在 K 上的**限制**. 因此

$$(f \mid K)(x) = f(x) \qquad (x \in K)$$

这个概念对样条函数来说是很有用的，这是因为每一个函数 $B_i^k \mid (t_j, t_{j+1})$ 是一个多项式(更准确地说，是一个多项式的限制). 当我们说函数 f_i 的集合**在集合 K 上线性无关**时，实际上指的是限制函数 $f_i \mid K$ 的集合在通常意义下线性无关.

现在考虑 B 样条 B_0^k，B_1^k，…，B_k^k. 当这些函数被限制在结点之间任何单个区间 (t_v, t_{v+1}) 上时，其结果是一组次数 $\leqslant k$ 的多项式. 一个令人惊讶而又有用的事实是这些限制函数构成了区间 (t_k, t_{k+1}) 上多项式空间 Π_k 的一个基.

引理 8(*B* 样条线性无关性引理) *B* 样条集合 $\{B_j^k, B_{j+1}^k, \cdots, B_{j+k}^k\}$ 在 (t_{k+j}, t_{k+j+1}) 上线性无关.

证明 首先考虑 $k=0$ 的情况. 引理断言 $\{B_j^0\}$ 在区间 (t_j, t_{j+1}) 上线性无关. 这显然是对的. 为了使用数学归纳法，令 $k \geqslant 1$ 并且假设对指标 $k-1$ 引理结论成立. 在此基础上，我们将对指标 k 证明引理. 令 $S = \sum_{i=0}^{k} c_{j+i} B_{j+i}^k$，并且假设 $S \mid (t_{k+j}, t_{k+j+1}) = 0$. 因为在 (t_{k+j}, t_{k+j+1}) 上 $B_{j+k+1}^{k-1} = 0$，$B_j^{k-1} = 0$，再根据(16)式，得到下面的等式

373

$$0 = S' \mid (t_{k+j}, t_{k+j+1}) = k\sum_{i=1}^{k}\frac{c_{j+i} - c_{j+i-1}}{t_{j+i+k} - t_{j+i}} B_{j+i}^{k-1} \mid (t_{k+j}, t_{k+j+1})$$

对 $\{B_{j+1}^{k-1}, B_{j+2}^{k-1}, \cdots, B_{j+k}^{k-1}\}$ 使用归纳假设知，在区间 (t_{k+j}, t_{k+j+1}) 上它们线性无关. 因此，(16)式中所有的系数一定是 0，从而我们有 $c_0 = c_1 = \cdots = c_k$. 如果我们把这个公共的值记为 λ，由引理 4 知在 (t_{k+j}, t_{k+j+1}) 上 $S(x) = \lambda$. (在引理 4 中，可以看出在区间 (t_{k+j}, t_{k+j+1}) 上仅有的

非零项是 B_j^k，B_{j+1}^k，…，B_{j+k}^k.）因为已经假设在区间$(t_{k+j},\ t_{k+j+1})$上 S 是 0，最终得到 $\lambda=0$. ■

引理 9（B 样条线性无关性引理） B 样条集合$\{B_{-k}^k,\ B_{-k+1}^k,\ \cdots,\ B_{n-1}^k\}$在$(t_0,\ t_n)$上线性无关.

证明 令 $S=\sum_{i=-k}^{n-1}c_iB_i^k$，并且假设 $S\mid(t_0,t_n)=0$. 在区间(t_0,t_1)上，仅有 $B_{-k}^k,B_{-k+1}^k,\cdots,B_0^k$ 是非零的，所以

$$0=S\mid(t_0,t_1)=\sum_{i=-k}^{0}c_iB_i^k\mid(t_0,t_1) \tag{17}$$

由引理 8 知集合$\{B_{-k}^k,\ B_{-k+1}^k,\ \cdots,\ B_0^k\}$在$(t_0,\ t_1)$上线性无关. 因此，根据(17)式，我们推出 $c_i=0$，$-k\leqslant i\leqslant 0$. 如果所有的 c_i 都是 0，我们就已获得所期望的结论. 否则，令 j 是第一个使得 $c_j\neq 0$ 的指标. 由前面的讨论知，$j\geqslant 1$，因此$(t_j,\ t_{j+1})\subseteqq(t_0,\ t_n)$. 对任意 $x\in(t_j,\ t_{j+1})$，得到下列矛盾

$$0=S(x)=\sum_{i=j}^{n-1}c_iB_i^k(x)=c_jB_j^k(x)\neq 0$$

因而，所有的 c_i 都是 0. ■

习题 6.5

1. 证明 $B_j^2(t_i)$的公式

$$B_j^2(t_i)=\left(\frac{t_i-t_{i-1}}{t_{i+1}-t_{i-1}}\right)\delta_{i,j+1}+\left(\frac{t_{i+1}-t_i}{t_{i+1}-t_{i-1}}\right)\delta_{i,j+2}$$

2. 证明：若 $t_m\leqslant x<t_{m+1}$，则

$$\sum_{i=-\infty}^{\infty}c_iB_i^k(x)=\sum_{i=m-k}^{m}c_iB_i^k(x)$$

3. 设 $h_i=t_{i+1}-t_i$. 证明：若

$$c_{i-1}h_{i-1}+c_{i-2}h_i=y_i(h_i+h_{i-1})\qquad(i\in \mathrm{Z})$$

则对所有 j，样条函数 $S=\sum_{i=-\infty}^{\infty}c_iB_i^2$ 具有插值性质 $S(t_j)=y_j$.

4. 假设结点选取所有的整数：$t_i=i(i\in \mathrm{Z})$. 证明：$B_i^k(x)=B_0^k(x-t_i)$.

5. 数一数在计算 $\sum_{i=-\infty}^{\infty}C_i^kB_i^k(x)$ 的算法中所包含的乘法、除法以及加法或减法运算的次数. 374

6. 证明关于样条导数的(16)式.

7. 证明：

$$\int_{-\infty}^{\infty}B_i^k(x)\mathrm{d}x=\frac{t_{i+k+1}-t_i}{k+1}$$

8. 证明：若对于所有 x 都有 $\sum_{i=-\infty}^{\infty}c_iB_i^k(x)=0$，则对于所有 i 都有 $c_i=0$.

9. 用下列公式定义 s 的函数：

$$U_i^k(s)=(t_{i+1}-s)(t_{i+2}-s)\cdots(t_{i+k}-s)$$

对于 $k=0$，设 $U_i^0(s)=1$. 证明

$$U_i^k(s)V_i^k(x)+U_{i-1}^k(s)[1-V_i^k(x)]=(x-s)U_i^{k-1}(s)$$

10. (续)证明：

$$\sum_{i=-\infty}^{\infty} U_i^k(s)B_i^k(x) = (x-s)\sum_{i=-\infty}^{\infty} U_i^{k-1}(s)B_i^{k-1}(x)$$

11. (续)证明 **Marsden 恒等式**

$$\sum_{i=-\infty}^{\infty} U_i^k(s)B_i^k(x) = (x-s)^k$$

12. (续)证明：每一个次数 $\leqslant k$ 的多项式都可以表示为 $\sum_{i=-\infty}^{\infty} c_i B_i^k$ 的形式.

13. 利用课本中的记号，证明 B_i^2 可由下列公式给出：

$$B_i^2(x) = \begin{cases} V_i^2 V_i^1 & x \in [t_i, t_{i+1}) \\ V_i^k - V_{i+1}^1(V_i^2 - 1 + V_{i+1}^2) & x \in [t_{i+1}, t_{i+2}) \\ (1 - V_{i+1}^2)(1 - V_{i+2}^1) & x \in [t_{i+2}, t_{i+3}) \\ 0 & \text{其他} \end{cases}$$

14. 验证下列两个等式：

$$\frac{\mathrm{d}}{\mathrm{d}x}B_i^1 = \frac{B_i^0}{t_{i+1} - t_i} - \frac{B_{i+1}^0}{t_{i+2} - t_{i+1}}$$

$$\frac{\mathrm{d}}{\mathrm{d}x}B_i^2 = \frac{2B_i^1}{t_{i+2} - t_i} - \frac{2B_{i+1}^1}{t_{i+3} - t_{i+1}}$$

15. 证明

$$\sup_{-\infty<x<\infty} \left| \sum_{i=-\infty}^{\infty} c_i B_i^k(x) \right| \leqslant \sup_{-\infty<i<\infty} |c_i|$$

16. 证明：若 $\sup_i |t_{i+1} - t_i| \leqslant m$，则

$$\int_{-\infty}^{\infty} \left| \sum_{i=-\infty}^{\infty} c_i B_i^k(x) \right| \mathrm{d}x \leqslant m \sum_{i=-\infty}^{\infty} |c_i|$$

17. 求出下面表达式的一个上界：

375

$$\int_{-\infty}^{\infty} \left[\sum_{i=-\infty}^{\infty} c_i B_i^k(x) \right]^2 \mathrm{d}x$$

18. 对于 $k \geqslant 3$，证明

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2} \sum_{i=-\infty}^{\infty} c_i B_i^k(x) = k(k-1) \sum_{i=-\infty}^{\infty} \left(\frac{c_i - c_{i-1}}{t_{i+k} - t_i} - \frac{c_{i-1} - c_{i-2}}{t_{i+k-1} - t_{i-1}} \right) \frac{B_i^{k-2}(x)}{t_{i+k-1} - t_i}$$

19. 证明：若 $a_i = k^{-1}(t_{i+1} + t_{i+2} + \cdots + t_{i+k})$，则

$$\sum_{i=-\infty}^{\infty} a_i B_i^k(x) = x \qquad (k \geqslant 1)$$

20. 给出一个样条函数 $\sum_{i=-\infty}^{\infty} c_i B_i^k$ 的例子，使得它不是 0 函数但是在所有结点上取零.

21. 假设在常数序列 C_{m-k}^k，C_{m-k+1}^k，…，C_m^k 中，其中有两个相邻的元素是相同的. 证明样条函数

$$\sum_{i=-\infty}^{\infty} C_i^k B_i^k(x)$$

是区间 (t_m, t_{m+1}) 上一个次数小于 k 的多项式.

22. 利用引理 3 后面的算法，证明 $\sum_{i=-\infty}^{\infty} c_i B_i^k$ 是 (t_m, t_{m+1}) 上一个次数 $\leqslant k$ 的多项式.

23. 利用引理 3 后面的算法，给出引理 1 和引理 2 的新的证明.

24. 证明：在 (t_j, t_{j+1}) 上 $B_i^k(x) > 0$ 当且仅当 $j - k \leqslant i \leqslant j$.

25. 对于 $k=1$，除三个结点 t_i，t_{i+1}，t_{i+2}之外，证明引理 5 中等式(10)成立.

26. 求出下列公式中的 α 和 β

$$\text{support}\left(\prod_{i=0}^{r} B_i^k\right)=\bigcap_{i=0}^{r}\text{support}(B_i^k)=(\alpha,\beta)$$

27. 使得$\{B_j^k, B_{j+1}^k, \cdots, B_{n+k+j-1}^k\}$在其上线性无关的最小区间是什么？

28. 证明

$$\sum_{i=0}^{n}B_i^k(x)=1 \qquad (t_k\leqslant x\leqslant t_{k+n})$$

29. 证明

$$\sum_{i=0}^{n}B_i^k(x)>0 \qquad (t_1<x<t_{n+k+1})$$

计算机习题 6.5

1. 编写并测试一个子程序用于计算区间$[a,b]$上 $g(x)=\int_a^x f(t)\mathrm{d}t$ 的近似值．函数 f 由用户在一个单独的函数子程序中提供.所使用的方法是：首先利用 n 个等距结点，用自然三次样条 S 在$[a,b]$上的插值 f，然后再用 $g(x)\approx\int_a^x S(t)\mathrm{d}t$.用户将具体指定 a,b 和 n.

2. 指定结点 t_1，t_2，…，t_{n+k+1}，以及 n 和 k．给定系数 c_1，c_2，…，c_n，并且记 $f(x)=\sum_{i=1}^{n}c_iB_i^k(x)$．编写一个子程序，对于任意实数 x 用该程序给出 $f(x)$的值.

3.（难题）设结点是全体整数集合，对于 $1\leqslant k\leqslant 100$，编写一个程序用于计算 B_0^k 的上确界范数．一个短程序就够了．测试情况：对于 $k=1$，5，10，其值分别是 1，0.55，0.410 963.

376

6.6 B 样条：应用

在本节中，我们仍采用上一节的记号，给定结点的初始序列为

$$\cdots<t_{-2}<t_{-1}<t_0<t_1<t_2<\cdots \tag{1}$$

在这些结点上，定义 $B_i^k(x)$是一组 B 样条，我们要把这些 B 样条与 6.4 节中最初引入的样条函数联系起来．在那里，我们考虑的函数是连续函数类 C^{k-1}中的整体连续函数，它是 n 个区间$[t_0, t_1]$，$[t_1, t_2]$，…，$[t_{n-1}, t_n]$上次数$\leqslant k$ 的分段多项式．用 $\mathcal{S}_n^k$ 表示所有这样的样条函数族．因为开始时结点是固定的，所以这个记号没有显示出它们．我们认为 $\mathcal{S}_n^k$ 中样条函数具有定义域$[t_0, t_n]$．那么这些函数与 B 样条 B_i^k 有怎样的关系呢？这里我们始终假设 $n\geqslant 1$.

6.6.1 空间 $\mathcal{S}_n^k$ 的基

我们把函数 B_i^k 限制在区间$[t_0, t_n]$上使得这些函数的定义域相同．所限制的函数记为 $B_i^k\mid[t_0, t_n]$.

定理 1（空间 $\mathcal{S}_n^k$ 的基定理） 空间 $\mathcal{S}_n^k$ 的一个基是

$$\{B_i^k\mid[t_0,t_n]:-k\leqslant i\leqslant n-1\} \tag{2}$$

从而，$\mathcal{S}_n^k$ 的维数是 $k+n$.

证明 首先，因为(2)式中的函数都是(1)式中给定的结点序列的 k 次样条函数，所以显然

它们属于 $\mathcal{S}_n^k$.

其次，我们证明 $\mathcal{S}_n^k$ 的维数不超过 $k+n$. 这一点可以由 $\mathcal{S}_n^k$ 是用 $k+n$ 个函数所生成来说明. 事实上，$\mathcal{S}_n^k$ 中的每个元素可表示成

$$S(x) = \sum_{i=0}^{k} a_i x^i + \sum_{i=1}^{n-1} b_i (x - t_i)_+^k \tag{3}$$

这个等式中用到以下截断幂函数

$$(x-t_i)_+^k = \begin{cases}(x-t_i)^k & 若\ x \geqslant t_i \\ 0 & 若\ x < t_i\end{cases}$$

为证明(3)式，我们从区间$[t_0, t_1]$开始，所有这些截断幂函数在该区间上都是 0. 在$[t_0, t_1]$上，$S(x)$是一个 k 次多项式，记为 p_0，从而有 $p_0(x) = \sum_{i=0}^{k} a_i x^i$，这样就确定了所有的系数 a_i. 在区间$[t_1, t_2]$上，$S(x)$是另一多项式，记为 p_1. 根据在点 t_1 的连续性，我们有

377

$$(p_1 - p_0)^{(r)}(t_1) = 0 \qquad (0 \leqslant r \leqslant k-1)$$

由于 $p_1 - p_0$ 的次数不超过 k 次，所以我们断定有 b_1 使得$(p_1-p_0)(x)=b_1(x-x_1)^k$，因此，有表达式

$$S(x) = \sum_{i=0}^{k} a_i x^i + b_1(x-t_1)_+^k \qquad (t_0 \leqslant x \leqslant t_2)$$

对 $t_2, t_3, \cdots, t_{n-1}$ 重复上面的讨论，可推导出(3)式中的其他项.

最后，应用 6.5 节中的引理 9 可推出(2)式中的函数组线性无关. 因而它是 $\mathcal{S}_n^k$ 的基. ■

定理 1 的证明过程表明函数组

$$1, x, x^2, \cdots, x^k, (x-t_1)_+^k, (x-t_2)_+^k, \cdots, (x-t_{n-1})_+^k$$

构成 $\mathcal{S}_n^k$ 的基. 因为它具有非常坏的条件，所以在把这个基用于数值计算时要谨慎. 取而代之，应该使用定理中所给出的 B 样条基.

6.6.2 插值矩阵

样条函数可用于结点以外点上的插值. 给定结点集：$x_1 < x_2 < \cdots < x_n$，我们希望用形如 $\sum_{j=1}^{n} c_j B_j^k$ 的样条函数插值结点上任意给定的数据. 为此，下列给出的**插值矩阵** A

$$A_{ij} = B_j^k(x_i) \qquad (1 \leqslant i, j \leqslant n) \tag{4}$$

必须是非奇异的. Schoenberg 和 Whitney 给出的一个漂亮定理揭示了这个非奇异性的充分必要条件：A 的主对角线上不含零元.

我们依据 de Boor[1976 以及私人书信]所给出的这个结果的证明，下面分为几种情况来考虑.

引理 1(插值矩阵引理 1) 若(4)式中的矩阵非奇异，则 $A_{ii} \neq 0$，$1 \leqslant i \leqslant n$.

证明 假设对某个 r 使得 $A_{rr}=0$，则由 6.5 节中引理 2 知 $B_r^k(x_r)=0$，$x_r \notin (t_r, t_{r+k+1})$. 首先，假设 $x_r \leqslant t_r$. 如果 $i \leqslant r \leqslant j$，那么 $x_i \leqslant x_r \leqslant t_r \leqslant t_j$，并且 x_i 不在 B_j 的支撑中. 因而，

$A_{ij}=B_j^k(x_i)=0$. 因为对于 $j=r, r+1, \cdots, n$ 都有 $A_{ij}=0$，所以 A 中前 r 行可以看成$\mathbb{R}^{r-1}$中的向量. 因此这 r 行线性相关，从而 A 奇异.

另一种情况，$x_r \geqslant t_{r+k+1}$. 如果 $i \geqslant r \geqslant j$，那么 $x_i \geqslant x_r \geqslant t_{r+k+1} \geqslant t_{j+k+1}$ 并且 $A_{ij}=B_j^k(x_i)=0$. 因为其第 $r, r+1, \cdots, n$ 项分量都是零，所以前 r 列向量线性相关. 再次得知 A 是奇异的. [378]

■

引理 2(插值矩阵引理 2) 若 $k=1$ 并且对于 $1 \leqslant i \leqslant n$ 有 $t_i < x_i < t_{i+2}$，则 A 非奇异.

证明 对 n 作数学归纳法. 若 $n=1$，则 A 是一个 1×1 矩阵并且有唯一元 $A_{11}=B_1^1(x_1) \neq 0$.

现假设结点个数小于 n 时引理 2 已被证明. 下面考虑 n 个结点的情况，这里 $n \geqslant 2$. 如果有一个指标 r 使得 $1 \leqslant r \leqslant n-1$ 并且 $x_r \leqslant t_{r+1}$，那么对于满足 $i \leqslant r < j$ 的任意一对指标(i, j)，我们有 $x_i \leqslant x_r \leqslant t_{r+1} \leqslant t_j$ 且 $A_{ij}=B_j^1(x_i)=0$. 因此，矩阵 A 有下列形式

$$A=\begin{bmatrix} C & 0 \\ E & D \end{bmatrix}$$

其中 C 是 $r \times r$ 矩阵，D 是$(n-r) \times (n-r)$矩阵. 矩阵 A 可逆当且仅当 C 和 D 可逆(见习题 6.1.11.)这时 C 和 D 是低阶矩阵(因为 $r<n$ 且 $n-r<n$)，否则 C 和 D 恰与 A 一致. 因此，由归纳假设知它们可逆.

对于 $2 \leqslant r \leqslant n$ 范围内的某些指标 r，如果 $x_r \geqslant t_{r+1}$，可进行同样的讨论. 这时，如果 $j<r \leqslant i$，那么 $A_{ij}=0$. A 的结构是

$$A=\begin{bmatrix} C & E \\ 0 & D \end{bmatrix}$$

其中 C 是$(r-1) \times (r-1)$矩阵，D 是$(n-r+1) \times (n-r+1)$矩阵.

仅剩的情况是当 $1 \leqslant i \leqslant n-1$ 时 $x_i > t_{i+1}$ 以及当 $2 \leqslant i \leqslant n$ 时 $x_i < t_{i+1}$. 而只有在 $n=1$(已经讨论过)或者 $n=2$ 时才会出现这种情况. 这时有 $n=2$，$k=1$，$x_1 > t_2$ 以及 $x_2 < t_3$. 因此 $t_2 < x_1 < < x_2 < t_3$. 在区间(t_2, t_3)上，有 $B_1^1(x)+B_2^1(x)=1$. 从而矩阵 A 形如

$$\begin{bmatrix} B_1^1(x_1) & B_2^1(x_1) \\ B_1^1(x_2) & B_2^1(x_2) \end{bmatrix}=\begin{bmatrix} \lambda & 1-\lambda \\ \mu & 1-\mu \end{bmatrix}$$

其行列式为 $\lambda-\mu \equiv B_1^1(x_1)-B_1^1(x_2)>0$. ■

接下来是 Schoenberg-Whitney 定理.

定理 2(Schoenberg-Whitney 定理) 设 $x_1 < x_2 < \cdots < x_n$. 已知 $A_{ij}=B_j^k(x_i)$，则矩阵 A 非奇异的充分必要条件是其主对角线上不含零元.

证明 必要性已由引理 1 给出. 因为 B 样条的支撑不相交，即 $B_j^0(x_i)=\delta_{ij}$，并且条件 $B_i^0(x_i) \neq 0$ 等价于 $t_i \leqslant x_i < t_{i+1}$，所以当 $k=0$ 时充分性显然成立.

当 $k=1$ 时，充分性已由引理 2 给出. 现在对 k 作数学归纳法. 假设对于次数小于 k 的 B 样条函数定理结论成立. 在此假设下，我们证明 k 次样条的情况. 由引理 2 知，这里 $k \geqslant 2$. 我们的证明还需要对 n 做数学归纳法，这一点类似于引理 2 的证明，但是这里的证明不太正式，现将证明的要点叙述如下：如果某个 $r \in \{1, 2, \cdots, n-1\}$ 使得 $x_r \leqslant t_{r+1}$ 或者如果某个 $r \in \{2, 3, \cdots, n\}$ 使得 $x_r \geqslant t_{r+k}$，那么就像引理 2 中那样，矩阵 A 具有分块结构，再把归纳假设用于 [379]

A 的两个子矩阵，可知矩阵 A 非奇异，因此我们可以假设

$$t_{i+1} < x_i < x_{i+1} < t_{i+k+1} \qquad (1 \leqslant i \leqslant n-1) \tag{5}$$

如果 A 奇异，那么存在 $u \neq 0$ 使得 $Au=0$. 令 $f = \sum_{j=1}^{n} u_j B_j^k$. 注意到在 $n+2$ 个点 t_1，x_1，x_2，…，x_n，t_{n+k+1} 上 f 是 0. 因为 $k \geqslant 2$ 时 f' 存在且连续. 因此可利用罗尔定理，我们推出存在 f' 的 $n+1$ 个零点，记为 ξ_i，整理如下：

$$t_1 < \xi_1 < x_1 < \xi_2 < x_2 < \cdots < x_{n-1} < \xi_n < x_n < \xi_{n+1} < t_{n+k+1}$$

如 6.5 节中(16) 式所示，f' 形如 $\sum_{j=1}^{n+1} v_j B_j^{k-1}$. 我们还有

$$\sum_{j=1}^{n+1} v_j B_j^{k-1}(\xi_i) = f'(\xi_i) = 0 \qquad (1 \leqslant i \leqslant n+1) \tag{6}$$

利用(5)式，得知 ξ_i 属于 B_i^{k-1} 的支撑，$2 \leqslant i \leqslant n$.

现在出现了几种情形. 情形 1：假设 $\xi_1 < t_{k+1}$ 并且 $\xi_{n+1} > t_{n+1}$. 那么 ξ_1 在 B_1^{k-1} 的支撑中且 ξ_{n+1} 在 B_{n+1}^{k-1} 的支撑中. 根据归纳假设，$(n+1)\times(n+1)$矩阵$(B_j^{k-1}(\xi_i))$非奇异. 因此(6)式中的系数 v_j 必须都是 0. 如同 6.5 节中(16)式一样，v_j 的实际表达式是

$$v_j = k(u_j - u_{j-1})/(t_{j+k} - t_j)$$

如果选取 $u_0 = u_{n+1} = 0$，对于 $1 \leqslant j \leqslant n+1$ 我们可以使用上述公式. 而系数 v_j 变为零必然有

$$u_1 - u_0 = u_2 - u_1 = \cdots = u_{n+1} - u_n = 0$$

由 $u_0 = u_{n+1} = 0$ 知 $u_i = 0$，$1 \leqslant i \leqslant n$.

情形 2：假设 $\xi_1 \geqslant t_{k+1}$ 并且 $\xi_{n+1} > t_{n+1}$. 因为 ξ_i 不在 B_1^{k-1} 的支撑中，所以对于 $1 \leqslant i \leqslant n+1$ 有 $B_1^{k-1}(\xi_i)=0$. 这时，由(6)式可推出

$$\sum_{j=2}^{n+1} v_j B_j^{k-1}(\xi_i) = 0 \qquad (2 \leqslant i \leqslant n+1)$$

但是当 $2 \leqslant i \leqslant n+1$ 时，ξ_i 在 B_i^{k-1} 的支撑中，故根据归纳假设，对于 $2 \leqslant i \leqslant n+1$，我们有 $v_j = 0$. 如前所示，这就导致

380

$$u_2 - u_1 = u_3 - u_2 = \cdots = u_{n+1} - u_n = 0$$

且由 $u_{n+1} = 0$ 知 $u_i = 0$，$1 \leqslant i \leqslant n$.

情形 3：假设 $\xi_1 < t_{k+1}$ 并且 $\xi_n \leqslant t_{n+1}$. 与情形 2 相似，当 $1 \leqslant i \leqslant n$ 时，ξ_i 在 B_i^{k-1} 的支撑中. 经与前面同样的讨论知 $u_i = 0$，$1 \leqslant i \leqslant n$.

情形 4：假设 $\xi_1 \geqslant t_{k+1}$ 并且 $t_{n+1} \geqslant \xi_{n+1}$. 对于 $1 \leqslant i \leqslant n+1$，我们有 $B_1^{k-1}(\xi_i)=0$ 以及 $B_{n+1}^{k-1}(\xi_i)=0$. 由(6)式得到

$$\sum_{j=2}^{n} v_j B_j^{k-1}(\xi_i) = 0 \qquad (2 \leqslant i \leqslant n)$$

但是 $2 \leqslant i \leqslant n$ 时，ξ_i 在 B_i^{k-1} 的支撑中，同理可知

$$u_2 - u_1 = u_3 - u_2 = \cdots = u_n - u_{n-1} = 0$$

这表明所有的 u_j 都相同，记为 λ. 因此 $f = \lambda \sum_{j=1}^{n} B_j^k$. 因为 $B_j^k \geqslant 0$ 且 $B_1^k(x_1) > 0$，所以等式 $f(x_1)=0$ 表明 $\lambda = 0$. ■

例 1 设结点为全体整数. 如果我们希望用 B_1^2, B_2^2, …, B_5^2 的线性组合作为插值函数, 那么可以使用结点集{3.1, 3.5, 3.6, 6.1, 6.6}吗?

解 这时由 Schoenberg-Whitney 定理给出的的条件可以写为

$$i < x_i < i+3 \qquad (1 \leqslant i \leqslant 5)$$

因而容易证明给定的结点满足这个条件. ■

6.6.3 存在性

由定理 1 可知样条空间 $\mathcal{S}_n^k$ 的维数是 $n+k$. 构成 $\mathcal{S}_n^k$ 的函数定义域是$[t_0, t_n]$. 如果在$[t_0, t_n]$中选定结点 x_1, x_2, …, x_{n+k}, 为了能用 $\mathcal{S}_n^k$ 插值, 这些结点必须满足什么条件?

定理 3(*B 样条插值定理*) *若*$[t_0, t_n]$*中的结点* x_1, x_2, …, x_{n+k}*满足*

$$t_{i-k-1} < x_i < t_i \qquad (1 \leqslant i \leqslant n+k)$$

则能用样条空间 $\mathcal{S}_n^k$ *插值这组结点上任意一组数据.*

证明 由定理 1 知, 函数集 $B_j^k \mid [t_0, t_n]$是 $\mathcal{S}_n^k$ 的一组基, 其中$-k \leqslant j \leqslant n-1$. 对于$-k \leqslant i \leqslant n-1$, 重新标注结点 $y_i = x_{i+1+k}$那么 $y_i \in \operatorname{supp}(B_i^k)$, 并且根据 Schoenberg-Whitney 定理知, 矩阵 $B_j^k(y_i)$非奇异. ■ 381

现在我们回到利用空间 $\mathcal{S}_n^k$ 对结点 x_1, x_2, …, x_n 的插值问题. 插值矩阵是 $A_{ij} = B_j^k(x_i)$由(4)式给出. 如果我们假设 $B_i^k(x_i) \neq 0$, $0 \leqslant i \leqslant n$ (这与假设 $t_i < x_i < t_{i+k+1}$ 相同), 由 Schoenberg-Whitney 定理知该矩阵非奇异. 因而可采用插值函数

$$S(x) = \sum_{j=1}^{n} c_j B_j^k(x)$$

这个等式中的系数可通过求解下列线性方程组得到:

$$\sum_{j=1}^{n} B_j^k(x_i) c_j = f(x_i) \qquad (1 \leqslant i \leqslant n) \tag{7}$$

这个插值问题的解未必是唯一的, 我们将通过考察在结点上插值的特殊情况来发现这种不唯一性. 假设 $x_i = t_i$. 有关这些结点的插值问题需要下列方程组的解:

$$\sum_{j=-\infty}^{\infty} c_j B_j^k(t_i) = f(t_i) \qquad (1 \leqslant i \leqslant n) \tag{8}$$

$k=0$ 是最简单的情况, 利用 $B_j^0(t_i) = \delta_{ij}$, 问题就立刻可解. 因此, 对于 $1 \leqslant j \leqslant n$, 有 $c_j = f(t_j)$. 因为 $B_j^1(t_i) = \delta_{i-1,j}$, 所以 $k=1$ 的情况同样可解. 因而, $c_{i-1} = f(t_i)$, $1 \leqslant i \leqslant n$.

当 $k=2$ 时, 求解这个插值问题要用到下面的事实(见习题 6.5.1):

$$B_j^2(t_i) = \left(\frac{t_i - t_{i-1}}{t_{i+1} - t_{i-1}}\right)\delta_{i-1,j} + \left(\frac{t_{i+1} - t_i}{t_{i+1} - t_{i-1}}\right)\delta_{i-2,j} \tag{9}$$

等式(9)直接表明(8)式中的每个线性方程只含两个未知量. 当然(8)式现在可改作下列形式:

$$\left(\frac{t_{i+1} - t_i}{t_{i+1} - t_{i-1}}\right)c_{i-2} + \left(\frac{t_i - t_{i-1}}{t_{i+1} - t_{i-1}}\right)c_{i-1} = f(t_i) \qquad (1 \leqslant i \leqslant n) \tag{10}$$

这个方程组含有 n 个方程及 $n+1$ 个未知量. 我们可以指定 c_{-1}的任意值, 并利用(10)式递归地计算出 c_0, c_1, …, c_{n-1}. 从而, 这个插值问题有很多解.

如果用同样的方式处理 $k=3$ 的情况, 那么我们就得到 n 个方程及 $n+2$ 个未知量. $k=3$

时方程组(8)是：

$$c_{i-3}B^3_{i-3}(t_i)+c_{i-2}B^3_{i-2}(t_i)+c_{i-1}B^3_{i-1}(t_i)=f(t_i)\qquad(1\leqslant i\leqslant n)\tag{11}$$

382 利用6.5节中(1)式可以计算出 $B^3_j(t_i)$ 的值. 其结果是

$$B^3_{i-3}(t_i)=\frac{(t_{i+1}-t_i)^2}{(t_{i+1}-t_{i-2})(t_{i+1}-t_{i-1})}$$

$$B^3_{i-2}(t_i)=\frac{(t_{i+2}-t_i)(t_i-t_{i-1})}{(t_{i+2}-t_{i-1})(t_{i+1}-t_{i-1})}+\frac{(t_i-t_{i-2})(t_{i+1}-t_i)}{(t_{i+1}-t_{i-2})(t_{i+1}-t_{i-1})}$$

$$B^3_{i-1}(t_i)=\frac{(t_i-t_{i-1})^2}{(t_{i+2}-t_{i-1})(t_{i+1}-t_{i-1})}$$

正像 $k=2$ 的情况那样，可以很容易地求解这个方程组；即指定 c_{-1} 和 c_{-2} 的任意值，利用(11)式可递归地计算出 c_0，c_1，…，c_{n-1}. 因为没有充分利用这两个额外参数 c_{-1} 和 c_{-2}，所以我们并不推荐方程组(11)的这种解法. 通常是在端点 t_1 和 t_n——即我们所关注区间的每个端点对样条增添附加条件. 例如，定义 $S''(t_1)=S''(t_n)=0$ 为自然样条条件.

对于自然样条插值，要求解的线性方程组是由(11)式以及习题6.6.8-10中给出的两个方程所组成的. 有关样条插值及其计算的安排，建议读者参阅 de Boor[1984].（在那本书里，B^k_j 表示 $k-1$ 次样条.）

6.6.4 非插值逼近方法

为解释非插值逼近方法，我们来了解 Schoenberg[1967]引入的一个简练的过程. 给定一个函数 f，我们用下式定义一个样条函数 Sf：

$$Sf=\sum_{i=-\infty}^{\infty}f(x_i)B^k_i\qquad x_i=\frac{1}{k}(t_{i+1}+\cdots+t_{i+k})\tag{12}$$

当 $k=0$ 时，令 $x_i=t_i$. 这时，(12)式就是前面已经讨论过的我们所熟知的插值格式. $k=1$ 的情况也与此类似. 但是对于更大的 k 值，(12)式所给出的样条函数 Sf 在任意指定的结点集上不插值 f.（这个算子称为**拟插值算子**.）这个逼近格式显著的性质是：

1. 若 f 是一个线性函数，则 $Sf=f$.
2. 对于任何线性函数 ℓ，$Sf-\ell$ 与 $f-\ell$ 仅存在符号差异.
3. 若 $f\geqslant0$，则 $Sf\geqslant0$.
4. 若 $|f|\leqslant M$，则 $|Sf|\leqslant M$.
5. S 是一个线性算子：$S(\alpha f+\beta g)=\alpha Sf+\beta Sg$.

对此感兴趣的读者可参阅 Marsden[1790]或者 Schoenberg[1967].

接下来的任务是讨论连续函数是否可以用样条函数逼近到任意精度. 在这种情况下，当为了提高精度而增加结点个数时，我们希望次数 k 保持不变. 现在所面临的问题是增加结点密度能否达到预期的精度；也就是说要对样条函数寻找一个类似魏尔斯特拉斯逼近的定理.

383 和前面一样，给定结点集

$$\cdots<t_{-2}<t_{-1}<t_0<t_1<t_2<\cdots$$

区间 $[t_0, t_n]$ 也记为 $[a, b]$，假设 f 是 $[a, b]$ 上一个已知函数. 函数 f 的扩张如图6-10所示. 因此，如果 f 在 $[a, b]$ 上连续，那么它的扩张也连续.

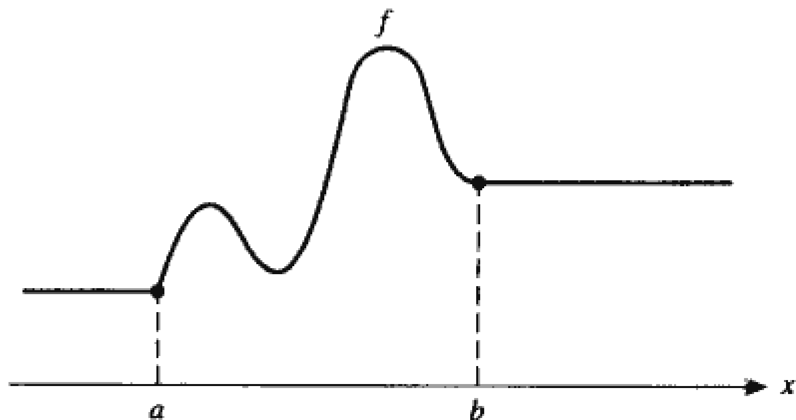


图 6-10　f 的扩张

不管 f 连续还是不连续，我们用下列等式来定义它的**连续模**：

$$\omega(f;\delta)=\max_{|s-t|\leqslant\delta}|f(s)-f(t)|$$

若 f 是 $[a,b]$ 上的连续函数，则它一致连续．这意味着对任意 $\varepsilon>0$，存在一个 $\delta>0$ 使得对 $[a,b]$ 中所有的 s 和 t，

$$|s-t|<\delta\Rightarrow|f(s)-f(t)|<\varepsilon$$

因此，$\omega(f;\delta)\leqslant\varepsilon$．换言之，对于有界闭区间上的一个连续函数 f，当 δ 收敛到 0 时其连续模 $\omega(f;\delta)$ 也收敛到 0.

如果 f' 存在，连续并且满足 $|f'(x)|\leqslant M$，那么由中值定理知

$$|f(s)-f(t)|=|f'(\xi)||s-t|\leqslant M|s-t|$$

因此，$\omega(f;\delta)\leqslant M\delta$.

下面引入一个样条函数，它以一种简单的方法逼近 f．为此，我们选择

$$g=\sum_{i=-\infty}^{\infty}f(t_{i+2})B_i^k \tag{13}$$

借助于这个函数，我们可以证明下面的结果．

384

定理 4（样条函数逼近定理）　若 f 是 $[t_0,t_n]$ 上的一个函数，则(13)式中的样条函数 g 满足

$$\max_{t_0\leqslant x\leqslant t_n}|f(x)-g(x)|\leqslant k\omega(f;\delta)$$

其中 $\delta=\max\limits_{-k\leqslant i\leqslant n+1}|t_i-t_{i-1}|$ 并且 $\omega(f;\cdot)$ 是 f 的连续模.

证明　因为 $B_i^k\geqslant0$ 并且 $\sum\limits_{i=-\infty}^{\infty}B_i^k=1$. 因而对于(13)式中的 g，有

$$\begin{aligned}|g(x)-f(x)|&=\left|\sum_{i=-\infty}^{\infty}f(t_{i+2})B_i^k(x)-f(x)\sum_{i=-\infty}^{\infty}B_i^k(x)\right|\\&=\left|\sum_{i=-\infty}^{\infty}[f(t_{i+2})-f(x)]B_i^k(x)\right|\\&\leqslant\sum_{i=-\infty}^{\infty}|f(t_{i+2})-f(x)|B_i^k(x)\end{aligned}$$

设 $x\in[t_j,t_{j+1}]\subseteq[a,b]$，在区间 $[t_j,t_{j+1}]$ 上，仅有 $B_{j-k}^k,B_{j-k+1}^k,\cdots,B_j^k$ 起作用．因此，

$$|g(x)-f(x)|\leqslant\sum_{i=j-k}^{j}|f(t_{i+2})-f(x)|B_i^k(x)$$

$$\leqslant \max_{j-k\leqslant i\leqslant j} |f(t_{i+2})-f(x)|$$

对于 $j-k\leqslant i\leqslant j$ 中的 i，有

$$t_{i+2}-x\leqslant t_{j+2}-t_j=(t_{j+2}-t_{j+1})+(t_{j+1}-t_j)\leqslant 2\delta$$

$$x-t_{i+2}\leqslant t_{j+1}-t_{j-k+2}=(t_{j+1}-t_j)+\cdots+(t_{j-k+3}-t_{j-k+2})\leqslant k\delta$$

根据连续模的定义以及习题 6.6.20，则有

$$|f(t_{i+2})-f(x)|\leqslant \omega(f;k\delta)\leqslant k\omega(f;\delta)$$

■

6.6.5 函数到样条空间的距离

上面的结果也可以说成是从函数 f 到空间 $\mathcal{S}_n^k$ 的距离. 我们把函数 f 到一个赋范空间的子空间 G 的距离定义为：

$$\text{dist}(f,G)=\inf_{g\in G}\| f-g\|$$

385 假设我们使用下面所定义的范数：

$$\| f\| = \max_{a\leqslant x\leqslant b} |f(x)|$$

由定理 4 推得

$$\text{dist}(f,\mathcal{S}_n^k)\leqslant k\omega(f;\delta) \tag{14}$$

如果 f 连续，那么

$$\lim_{\delta\downarrow 0}\omega(f;\delta)=0$$

因此，当结点密度增加时，(14)式中的上界将趋于 0.

对于某些具有导数的函数，会有更多的结果.

定理 5(函数到样条空间距离的定理) 设 $r<k<n$，若 $f\in C^r[t_0, t_n]$，则(其中 δ 与样条函数逼近定理中一致)

$$\text{dist}(f,\mathcal{S}_n^k)\leqslant k^r\delta^r\| f^{(r)}\|$$

证明 设 g 是 $\mathcal{S}_n^k$ 中的任意元. 根据定理 4，我们有

$$\text{dist}(f,\mathcal{S}_n^k)=\text{dist}(f-g,\mathcal{S}_n^k)\leqslant k\omega(f-g;\delta)\leqslant k\delta\| f'-g'\|$$

当 g 取遍 $\mathcal{S}_n^k$ 中的元时，g' 也取遍 S_n^{k-1} 中的元. (见习题 6.6.13)因此，在上述不等式中对 g 取下确界，得到

$$\text{dist}(f,\mathcal{S}_n^k)\leqslant k\delta\,\text{dist}(f',\mathcal{S}_n^{k-1})$$

再重复上述讨论过程 $r-2$ 次后就产生

$$\begin{aligned}\text{dist}(f,\mathcal{S}_n^k)&\leqslant k^{r-1}\delta^{r-1}\text{dist}(f^{(r-1)},\mathcal{S}_n^{k+1-r})\\&\leqslant k^r\delta^{r-1}\omega(f^{(r-1)};\delta)\\&\leqslant k^r\delta^r\| f^{(r)}\|\end{aligned}$$

■

习题 6.6

1. 证明：出现在 B 样条插值中的矩阵 $B_j^k(x_i)$ 是带状的. 特别是，若 $B_j^k(x_j)\neq 0$ 并且对每一个 j 都有 $x_j<x_{j+1}$，则该矩阵的每一行及每一列最多有 $2k+1$ 个非零元.
2. 在(12)式中，设 $k=2$，并且 $f(x)=x$，证明：$Sf=f$.
3. 证明：

$$x^2=\sum_{i=-\infty}^{\infty} t_{i+1}t_{i+2}B_i^2(x)$$

4. 在(12)式中，设 $k=2$，并且 $f(x)=1$. 证明：$Sf=f$..
5. 证明 Schoenberg 逼近过程中的性质 5.

386

6. 在 $k=2$ 的情况下，证明 Schoenberg 逼近过程中的性质 1. 提示：习题 6.6.2，4－5 会有帮助.
7. 证明 Schoenberg 过程中的性质 3 和性质 4.
8. 证明：若 $S=\sum_{j=-\infty}^{\infty} c_j B_j^3$，则 $S''=\sum_{j=-\infty}^{\infty} e_j B_j^1$，其中

$$e_j=\frac{6}{t_{j+2}-t_j}\left(\frac{c_j-c_{j-1}}{t_{j+3}-t_j}-\frac{c_{j-1}-c_{j-2}}{t_{j+2}-t_{j-1}}\right)$$

9. (续)证明：若 $S=\sum_{j=-\infty}^{\infty} c_j B_j^3$，则 $S''(t_i)=e_{i-1}$，其中 e_j 已在上题中定义.
10. (续) 证明：函数 f 的自然三次样条插值是 $\sum_{i=-2}^{n-1} c_j B_j^3$，其中系数满足课本中的(11) 式，并且对应于 $i=1$ 和 $i=n$ 的这些自然端点条件：

$$(t_{i+2}-t_{i-1})c_{i-3}-(t_{i+2}+t_{i+1}-t_{i-1}-t_{i-2})c_{i-2}+(t_{i+1}-t_{i-2})c_{i-1}=0$$

11. 设 C 和 D 是方阵，证明：

$$\begin{bmatrix} C & 0 \\ E & D \end{bmatrix}$$

是非奇异的当且仅当 C 和 D 是非奇异的.

12. 设 K 是 $\mathbb{R}$ 的任意子集. 证明：$\{B_1^k, \cdots, B_n^k\}$ 在 K 上线性无关的充分必要条件是对于 $1\leqslant i\leqslant n$ 集合 $K\cap(t_i, t_{i+k+1})$ 是非空的.
13. 利用定理 1 证明导数算子是从 $\mathcal{S}_n^k$ 到 $\mathcal{S}_n^{k-1}$ 上的满射.
14. 证明：6.5 节中的引理 8 可以作为 Schoenberg-Whitney 定理的一个推论.
15. 证明：6.5 节中的引理 9 可以作为 Schoenberg-Whitney 定理的一个推论.
16. 下面的叙述是 Schoenberg-Whitney 定理的一种正确说法吗？如果结点 x_i 不必有序，那么矩阵 $(B_j^k(x_i))$ 的非奇异性等价于论断：对于 $1\leqslant i\leqslant n$，每一个区间 (t_i, t_{i+k+1}) 至少包含一个结点.
17. 利用定理 4 证明过程中的方法，对 Schoenberg 算子证明一个类似的结果.
18. (续)对于等距结点的情况，改进上题中所得到的结果.
19. 证明在区间 $[t_0, t_n]$ 上仅有 B_i^k 序列的 $n+k$ 个元有非零值，再证明定理 1.
20. 证明：$\omega(f; k\delta)\leqslant k\omega(f; \delta)$.
21. 对所有的 k 值，证明 Schoenberg 过程中的性质 1. 由于线性性，只要证明当 $f(x)=1$ 或者 $f(x)=x$ 时一定有 $Sf=f$ 即可. 第一种情况很容易证明. 为了证明第二种情况，可利用 6.5 节中的(16)式，证明 $(Sf)'=1$.

计算机习题 6.6

设 $k=3$ 并选取结点是整数. 用 f_j 表示 $\mathcal{S}_n^k$ 中由幂函数和截断幂函数(像定理 1 的证明中一样)所构成的基，对于 $n=5$、10、15 和 20，估计矩阵 $(f_j(x_i))$ 的条件. 选取的 x_i 是 $[t_0, t_n]$ 中的等距点.

387

6.7 泰勒级数

本节比较简短，我们主要说明(并且强调)泰勒级数作为一种技巧在逼近过程中的多种用途. 当然，对于那些具有多阶连续导数的函数来说，泰勒定理是很有用的，然而，它在处理经验数据或者仅存在低阶导数的函数时并不是很有效. 对于那些可以应用泰勒定理的函数，千万

不要忽略了用泰勒多项式来有效地表示它们的可能性.

回顾前文知，如果函数 f 在区间$(c-\delta,\ c+\delta)$上有 $n+1$ 阶连续导数，那么

$$f(x)=p_n(x)+E_n(x)$$

其中 p_n 是次数$\leqslant n$ 的多项式，E_n 是**余项函数**，它们分别是

$$p_n(x)=\sum_{k=0}^{n}\frac{1}{k!}f^{(k)}(c)(x-c)^k$$

$$E_n(x)=\frac{1}{(n+1)!}f^{(n+1)}(\xi_x)(x-c)^{n+1}\qquad |\xi_x-c|<\delta$$

当 $c=0$ 时出现一个重要的特例，这就是**麦克劳林级数**.

当 $n\to\infty$时，通过分析 $E_n(x)$，再利用泰勒定理，我们可以得到很多重要函数的泰勒级数. 例如：

$$\cos x=\sum_{k=0}^{\infty}(-1)^k\frac{x^{2k}}{(2k)!}\qquad(-\infty<x<\infty)\tag{1}$$

$$\frac{1}{x}=\sum_{k=0}^{\infty}(-1)^k(x-1)^k\qquad(0<x<2)\tag{2}$$

上述过程中出现的级数是**幂级数**. 下面是有关幂级数的收敛性定理.

定理 1(幂级数收敛性定理) 对每一个幂级数

$$\sum_{k=0}^{\infty}a_k(x-c)^k$$

在区域$[0,\ \infty)$中都存在一个数 r，使得对于 $|x-c|>r$ 该级数发散；而对于 $|x-c|<r$ 这个级数收敛.

数 r (可以是$+\infty$)被称为级数的**收敛半径**. 通常用**比率检验法**可以求出它. (见习题 6.7.2～6.7.3.)(1)式中余弦函数的收敛半径是$+\infty$；(2)式中级数的收敛半径是$+1$.

388

下面的定理在应用方面很重要.

定理 2(收敛半径定理) 设级数 $\sum_{k=0}^{\infty}a_k(x-c)^k$ 的收敛半径是 r,则等式

$$f(x)=\sum_{k=0}^{\infty}a_k(x-c)^k$$

定义的函数在区间 $|x-c|<r$ 内连续可微. 而且,

$$f'(x)=\sum_{k=0}^{\infty}ka_k(x-c)^{k-1}$$

并且这个级数有收敛半径 r. 最后，若 $|b-c|<r$, $|x-c|<r$，则

$$\int_b^x f(t)\,\mathrm{d}t$$

可由 f 级数逐项积分得到，而逐项积分后所得到的级数也有收敛半径 r.

总之，该定理说明在幂级数的收敛区间内，可以对它进行逐项积分和逐项微分.

作为这个定理作用的一个例证，我们来考虑一个**高等超越函数——正弦积分**，它由下列公式定义

$$S(x)=\int_0^x \frac{\sin t}{t}\mathrm{d}t$$

对这个积分不可按照初等微积分的常用技巧来处理. 但是，我们可以如下进行：

$$\sin t=\sum_{k=0}^{\infty}(-1)^k\frac{t^{2k+1}}{(2k+1)!}$$

$$\frac{\sin t}{t}=\sum_{k=0}^{\infty}(-1)^k\frac{t^{2k}}{(2k+1)!}$$

$$\int_0^x\frac{\sin t}{t}\mathrm{d}t=\sum_{k=0}^{\infty}(-1)^k\frac{1}{(2k+1)!}\int_0^x t^{2k}\mathrm{d}t$$

$$S(x)=\sum_{k=0}^{\infty}(-1)^k\frac{x^{2k+1}}{(2k+1)!(2k+1)}$$

我们所得到的 $S(x)$ 级数对于较小的 x 值是快速收敛的. 例如，仅选择 10 项，$S(1)$ 的计算精度就可达到 20 位小数.（见习题 6.7.36.）

389

另一个类似的例子是

$$\int_0^x e^{t^2}\mathrm{d}t=\sum_{k=0}^{\infty}\frac{x^{2k+1}}{(2k+1)k!} \tag{3}$$

习题 6.7

1. 证明：$\sum_{k=0}^{\infty}a_kx^k$ 和 $\sum_{k=0}^{\infty}a_k(x-c)^k$ 有相同的收敛半径. 对于自然数 p，试问 $\sum_{k=0}^{\infty}a_kx^{k+p}$ 的收敛半径是什么？

2.（**比率检验法**）若 $\lim_{n\to\infty}|A_{n+1}/A_n|<1$，则 $\sum_{k=0}^{\infty}A_k$ 收敛. 用这个检验法证明(1)式中的余弦级数对所有的实数 x 收敛.

3.（**比率检验法，续**）若 $\lim_{n\to\infty}|A_{n+1}/A_n|>1$，则 $\sum_{k=0}^{\infty}A_k$ 发散. 用这个事实以及上一题，求出下列级数的收敛半径

$$\sum_{k=0}^{\infty}(-1)^k(x-1)^k$$

4. 求出下列级数的收敛半径：

$$\sum_{k=0}^{\infty}k!x^k$$

5. 若 $f(x)=\sum_{k=0}^{\infty}a_k(x-c)^k$ 并且收敛半径是 r，则 f 在区间 $|x-c|<r$ 内有任意阶导数. 此外，

$$f^{(n)}(x)=\sum_{k=n}^{\infty}\frac{a_kk!}{(k-n)!}(x-c)^{k-n}\qquad(|x-c|<r)$$

用收敛半径定理 2 来证明这个结论.

6. 求出下列积分的幂级数：

$$\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\mathrm{d}t$$

注：这个函数被称为**误差函数**，记为 $\mathrm{erf}(x)$，常用于统计学中.

7. 求出下列函数的幂级数：

$$f(x)=\int_0^x \frac{e^t-1}{t}dt$$

利用你得到的级数以三位有效数字计算 $f(1)$.(在此例中，要保留两位小数.)最后，证明你所舍去的所有项的和很小，因而不会影响答案. 注：函数 f 是应用数学中一个很重要的函数. 更多的信息可参阅 Abramowitz and Stegun[1964，第 5 章].

8. 证明：函数 $f(x)=\sum_{k=0}^{\infty}x^{2k}/(k!2^k)$ 是微分方程 $y'=xy$ 的解.

9. 设 $a_0=1$ 并且 $a_n=a_{n-1}/[2(n+1)]$，$n\geqslant 1$，那么 $\sum_{k=0}^{\infty}a_kx^k$ 的收敛半径是什么？

[390] 10. (续)设函数 $f(x)$ 的级数由上题给出. 试问 $f'(x)$ 的幂级数是什么？并给出系数所满足的递归关系.

11. 用比率检验法(见习题 6.7.2～6.7.3)，证明：若 $\lim_{n\to\infty}|a_n/a_{n+1}|$ 存在或者是 $+\infty$，则它是 $\sum_{k=0}^{\infty}a_k(x-c)^k$ 的收敛半径.

12. 设 $a_0=1$ 并且 $a_{n+1}=[2+(-1)^n]a_n, n\geqslant 1$. $\sum_{k=0}^{\infty}a_k(x-c)^k$ 的收敛半径是什么？

13. 如果 r 和 r' 分别是 $\sum_{k=0}^{\infty}a_kx^k$ 和 $\sum_{k=0}^{\infty}b_kx^k$ 的收敛半径，试问 $\sum_{k=0}^{\infty}(a_k+b_k)x^k$ 的收敛半径是什么？

14. 一个与**二重对数**有关的函数定义如下

$$f(x)=-\int_0^x \frac{\ln(1-t)}{t}dt \qquad (-\infty<x\leqslant 1)$$

求出 f 的麦克劳林级数并确定其收敛半径. 如何计算 $f(-2)$？$f(0.001)$ 又怎么样？

15. 通过(2)式中级数的积分求出 $\ln x$ 的一个级数.

16. 说明如何通过(1)式中级数的积分或微分这两种方法来求出 $\sin x$ 的级数.

17. 按下列步骤得到 $\tan^{-1}x$ 的一个幂级数：从 $(1+x)^{-1}=\sum_{k=0}^{\infty}(-x)^k$ 开始，用 x^2 替换 x，再对等式中的各项积分. 另一个过程是：计算 $\tan^{-1}x$ 的逐次导数并且求出泰勒级数，试比较上面这两个过程.

18. 评判下列解析过程并改正之：

$$\int_0^x \frac{1-\cos t}{t}dt=\int_0^x\left[\frac{1}{t}-\frac{\cos t}{t}\right]dt=\int_0^x\left[\frac{1}{t}-\sum_{k=0}^{\infty}\frac{(-1)^kt^{2k-1}}{(2k)!}\right]dt$$

$$=\ln x-\sum_{k=0}^{\infty}\frac{(-1)^kx^{2k}}{2k(2k)!}$$

提示：有多处错误.

19. 求出下列函数的麦克劳林级数：

$$f(x)=\int_0^x \frac{1-\cos t}{t}dt$$

注：这个函数有时被称为**余弦积分**并且记为 $\mathrm{Cin}(x)$. 但是该术语还没有标准化.

20. 求出**菲涅耳积分**的麦克劳林级数：

$$\varphi(x)=\int_0^x \sin t^2\,dt$$

21. (**互反函数**)假设 $f(x)=\sum_{k=0}^{\infty}a_kx^k$，并且 $a_0=1$. 那么，在 0 的某领域内，$1/f(x)$ 有定义. 假设互反函数的级数是 $\sum_{k=0}^{\infty}b_kx^k$，那么根据两个级数的乘积是 1，用递归关系确定 b_k.

[391] 22. (续)用上题的结果，求出 $x/(e^x-1)$ 的麦克劳林级数.

23. (续)证明：习题 6.7.21 中的系数 b_k 具有性质：$b_3=b_5=b_7=\cdots=0$. 注：系数 $k!\,b_k$ 称为**伯努利数**，常记

为 B_0，B_1，…，B_k.

24. (续)证明：习题 6.7.23 中伯努利数 B_k 具有性质：

$$\sum_{k=0}^{n-1}\binom{n}{k}B_k = 0 \qquad (n>1)$$

证明：若 $B_0=1$，这个等式可用来计算 B_1，B_2，…，并且有 $B_0=1$，$B_1=-1/2$，$B_2=1/6$，$B_3=0$，$B_4=-1/30$.

25. 证明：若 r 是级数 $\sum_{k=0}^{\infty}a_kx^k$ 的收敛半径，则导数级数 $\sum_{k=0}^{\infty}ka_kx^{k-1}$ 的收敛半径也是 r.

26. 证明：

$$\frac{x+1}{x-1} = 1-2\sum_{k=0}^{\infty}x^k \qquad (|x|<1)$$

27. 求出函数 $\csc x - 1/x$ 关于点 0 的幂级数.

28. 求出由下列等式定义的函数 f 的麦克劳林级数：

$$-\ln(1-x)f(x) = \frac{x}{1-x}$$

29. 将下列函数展开成 $(x-1)$ 的幂级数：

$$f(x) = \int_1^x \frac{e^t - e}{t-1}dt$$

并用所得级数计算 $f(0)$，精确到三位有效数字.

30. 建立公式(3).

31. 级数 $\sum_{k=0}^{\infty}k(2x)^k$ 的收敛半径是什么？

32. 如果 $c_0=1$ 并且 $c_n=(n/3)c_{n-1}$，$n\geqslant 1$，那么级数 $\sum_{k=0}^{\infty}c_kx^k$ 的收敛半径是什么？

33. 如果 $c_0=1$ 并且 $c_{n+1}=(3n+3)c_{n-1}/(n+5)$，$n\geqslant 1$，那么级数 $\sum_{k=0}^{\infty}c_kx^k$ 的收敛半径是什么？

34. 求出 $e^{\cos x}$ 关于 x 的幂级数中的前 3 项. 提示：令 $z=1-\cos x$.

35. 求出下列积分的一个幂级数：

$$\int_0^x e^{-t}dt$$

36. 证明：只需要正弦积分 $S(x)$ 的级数中的 10 项，就可以使 $S(1)$ 的计算精度达到 20 位小数. 那么当精度达到 25 位小数时，需要多少项？

计算机习题 6.7

编写一个程序或者子程序，对区间[−1，1]中的任意 x，计算具有 10 位小数位精度的正弦积分. 假设级数中的项数与 x 有关. 对区域以外的 x 程序将返回一个错误.

6.8 最佳逼近：最小二乘理论

最佳逼近的经典问题之一可叙述如下：已知区间$[a, b]$上的连续函数 f，对某一固定整数 n，找一个次数至多是 n 次的多项式 p，使其与 f 的偏差尽可能小. 偏差可用下列表达式度量 392

$$\max_{a\leqslant x\leqslant b}|f(x)-p(x)|$$

显而易见，这个问题完全不同于在某个给定结点集上简单地插值 f. 而且即使 f 能展开成泰勒级数，这个问题也不同于简单地截断 f 的泰勒级数. 事实上，这是一个**极值问题**，它的解一点也不明显.

6.8.1 存在性

一般的线性最佳逼近问题就包括刚才所描述的问题. 在这里，我们用 E 表示任意的赋范线性空间，并用 G 表示它的子空间. 对任一 $f\in E$，f 到 G 的距离定义为：

$$\text{dist}(f,G)=\inf_{g\in G}\|f-g\|$$

这个距离度量出我们用 G 中的一个元逼近向量 f 所能达到的绝对极小偏差. 如果 G 的一个元 g 有性质

$$\|f-g\|=\text{dist}(f,G)$$

那么 g 达到了这个极小偏差并且被称为 G 中 f 的**最佳逼近**. 因此，最佳逼近的意义与问题中所选择的范数有关.

在上面提到的经典问题中，赋范空间 E 是定义在 $[a,\ b]$ 上的全体连续函数空间 $C[a,\ b]$，范数是

$$\|f\|=\max_{a\leqslant x\leqslant b}|f(x)|\qquad f\in C[a,b]\tag{1}$$

子空间 G 则是全体次数 $\leqslant n$ 的多项式空间 Π_n，它是 $C[a,\ b]$ 的一个子空间. 后面 6.9 节将专门用来讨论这种特殊类型的逼近.

在一般最佳逼近问题中，焦点之一就是 f 的最佳逼近是否存在. 下面给出一个重要的存在性定理.

定理 1（*最佳逼近存在性定理*） *若 G 是赋范线性空间 E 中一个有限维子空间，则 E 的每一元在 G 中至少有一个最佳逼近.*

证明 设 f 是 E 的一个元，如果 g 是 f 最佳逼近的任一候选者，那么 g 与 G 的 0 元相比，有

$$\|f-g\|\leqslant\|f-0\|=\|f\|$$

从而我们可把搜索范围限制在下面集合中：

393

$$K=\{g\in G:\|g-f\|\leqslant\|f\|\}$$

K 是有界闭集. 因为 G 是有限维的，所以 K 是紧集. 因为函数 $g\mapsto\|f-g\|$ 是连续的，所以我们可以引用紧集上的连续实值函数能达到下确界这个定理. ■

最佳逼近一般不是唯一的. 下面的例子很容易说明这一点. 用函数 $g(x)=\lambda x$ 逼近区间 $[0,\ \pi/2]$ 上的函数 $f(x)=\cos x$，其中 λ 是我们选取的常数. 当采用(1)式中的范数时，几个最佳逼近如图 6-11 所示.

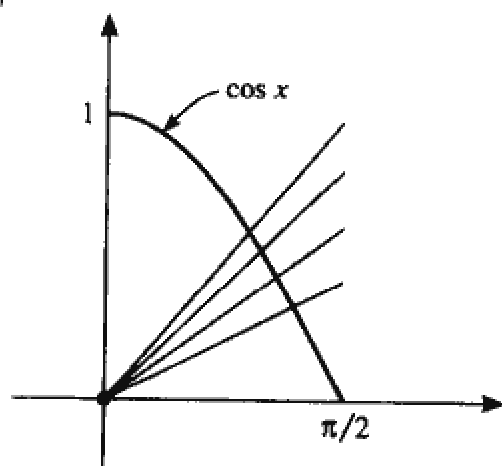


图 6-11 用 λx 逼近 $\cos x$

6.8.2 内积空间

对具体的 E，G 和 f 来说，要获得相应的最佳逼近是一件困难的事. 通常必须求解非线性方程组. 然而，有一种重要情况是我们只需要求解一些线性方程. 这就是当 E 是一个内积空

间时.

回忆一个实内积空间就是一个其中引入了内积和范数的线性空间 E.

公理 1(内积公理)

a. $\langle f, h\rangle=\langle h, f\rangle$.

b. $\langle f, \alpha h+\beta g\rangle=\alpha\langle f, h\rangle+\beta\langle f, g\rangle$.

c. 若 $f\neq 0$, 则 $\langle f, f\rangle>0$.

d. $\| f\| =\sqrt{\langle f, f\rangle}$.

一个重要的内积空间是$\mathbb{R}^n$, 具有内积

$$\langle x,y\rangle = \sum_{i=1}^{n} x_i y_i$$

另一个重要的内积空间是$[a, b]$上连续函数空间 $C_w[a, b]$, 具有

$$\langle f,g\rangle = \int_a^b f(x)g(x)w(x)\mathrm{d}x$$

394

其中 w 是一个取定的连续正函数. 在任一内积空间中, 如果$\langle f, g\rangle=0$, 我们记 $f\perp g$. 如果对于所有 $g\in G$ 都有 $f\perp g$, 我们记 $f\perp G$.

引理 1(内积空间性质引理) 在一个内积空间中, 我们有

1. $\left\langle \sum_{i=1}^{n} a_i f_i, g\right\rangle = \sum_{i=1}^{n} a_i\langle f_i, g\rangle$.
2. $\| f+g\|^2=\| f\|^2+2\langle f, g\rangle+\| g\|^2$.
3. 若 $f\perp g$, 则 $\| f+g\|^2=\| f\|^2+\| g\|^2$.
4. $|\langle f, g\rangle|\leqslant\| f\| \, \| g\|$.
5. $\| f+g\|^2+\| f-g\|^2=2\| f\|^2+2\| g\|^2$.

证明 根据公理 a 和 b, 用数学归纳法可以证明性质 1. 根据公理 a 和 b 以及 d 中的定义, 可得性质 2:

$$\begin{aligned}\| f+g\|^2 &= \langle f+g,f+g\rangle = \langle f,f\rangle+\langle f,g\rangle+\langle g,f\rangle+\langle g,g\rangle \\ &= \| f\|^2+2\langle f,g\rangle+\| g\|^2\end{aligned}$$

性质 3 称为**毕达哥拉斯法则**, 可由性质 2 立即推出. 性质 4 称为**施瓦茨不等式**. 为了证明它, 假设有元素对(f, g)使得

$$|\langle f,g\rangle| > \| f\| \, \| g\|$$

由于$\langle f, 0\rangle=0$, 显然 $g\neq 0$. 根据齐次性, 我们可以假设$\| g\|=1$, 使得$|\langle f, g\rangle|>\| f\|$. 再根据齐次性, 我们可假设$\langle f, g\rangle=1$, 从而$\| f\|<1$. 然而出现了矛盾:

$$0\leqslant \| f-g\|^2 = \| f\|^2-2\langle f,g\rangle+\| g\|^2 = \| f\|^2-1$$

根据性质 2, 直接计算即可得性质 5. ∎

一旦可以利用施瓦茨不等式, 范数的三角不等式就可以证明如下:

$$\begin{aligned}\| f+g\|^2 &= \| f\|^2+2\langle f,g\rangle+\| g\|^2 \\ &\leqslant \| f\|^2+2\| f\| \, \| g\|+\| g\|^2 \\ &= (\| f\|+\| g\|)^2\end{aligned}$$

定理2(刻画最佳逼近特性的定理) 设G是内积空间E的子空间. 对$f\in E$, $g\in G$, 下列性质等价：

1. g是G中f的一个最佳逼近.

395

2. $f-g\perp G$.

证明 如果$f-g\perp G$，那么根据毕达哥拉斯法则，对任一$h\in G$我们有

$$\| f-h \|^2 = \| (f-g)+(g-h) \|^2 = \| f-g \|^2 + \| g-h \|^2 \geqslant \| f-g \|^2$$

反之，假设g是f的一个最佳逼近. 设$h\in G$以及$\lambda>0$. 那么

$$\begin{aligned} 0 &\leqslant \| f-g+\lambda h \|^2 - \| f-g \|^2 \\ &= \| f-g \|^2 + 2\lambda\langle f-g,h\rangle + \lambda^2 \| h \|^2 - \| f-g \|^2 \\ &= \lambda\{2\langle f-g,h\rangle + \lambda \| h \|^2\} \end{aligned}$$

令$\lambda\downarrow 0$，得到$\langle f-g, h\rangle\geqslant 0$. $-h$也满足同样的不等式，从而有$\langle f-g, h\rangle\leqslant 0$. 因此，$\langle f-g, h\rangle=0$. 因为$h$是$G$中的任意元，所以$f-g\perp G$. ■

注意到证明所产生的一个结果就是特定条件下最佳逼近的唯一性.

6.8.3 正规方程

例1 根据定理2，用多项式$g(x)=c_1x+c_2x^3+c_3x^5$，确定区间$[-1, 1]$上$f(x)=\sin x$的最佳逼近. 使用范数

$$\| f \| = \left\{\int_{-1}^{1}[f(x)]^2\mathrm{d}x\right\}^{1/2}$$

解 最优函数g具有性质$f-g\perp G$，其中G是由$g_1(x)=x$, $g_2(x)=x^3$，以及$g_3(x)=x^5$生成的空间. 因而，需要$\langle g-f, g_i\rangle=0$，$1\leqslant i\leqslant 3$. 这些方程可写成：

$$c_1\langle g_1,g_i\rangle + c_2\langle g_2,g_i\rangle + c_3\langle g_3,g_i\rangle = \langle f,g_i\rangle \qquad (i=1,2,3)$$

并且该方程在这个问题中被称为**正规方程**. 给出其详细表达式，我们有

$$\begin{cases} c_1\int_{-1}^{1}x^2\mathrm{d}x + c_2\int_{-1}^{1}x^4\mathrm{d}x + c_3\int_{-1}^{1}x^6\mathrm{d}x = \int_{-1}^{1}x\cdot\sin x\mathrm{d}x \\ c_1\int_{-1}^{1}x^4\mathrm{d}x + c_2\int_{-1}^{1}x^6\mathrm{d}x + c_3\int_{-1}^{1}x^8\mathrm{d}x = \int_{-1}^{1}x^3\cdot\sin x\mathrm{d}x \\ c_1\int_{-1}^{1}x^6\mathrm{d}x + c_2\int_{-1}^{1}x^8\mathrm{d}x + c_3\int_{-1}^{1}x^{10}\mathrm{d}x = \int_{-1}^{1}x^5\cdot\sin x\mathrm{d}x \end{cases}$$

计算出所有积分后，我们要求解下面这个具有3×3系数矩阵的线性方程组：

396

$$\begin{bmatrix} \frac{1}{3} & \frac{1}{5} & \frac{1}{7} \\ \frac{1}{5} & \frac{1}{7} & \frac{1}{9} \\ \frac{1}{7} & \frac{1}{9} & \frac{1}{11} \end{bmatrix}\begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} \alpha-\beta \\ -3\alpha+5\beta \\ 65\alpha-101\beta \end{bmatrix}$$

其中$\alpha=\sin 1$以及$\beta=\cos 1$. 用α和β的具体数值，我们求得$c_1\approx -0.999\,98$, $c_2\approx -0.166\,52$, $c_3=-0.008\,02$. 回顾5.3节中对正规方程以及与其相关的数值微分的讨论. 这个系数矩阵也是2.3节中所见病态希尔伯特矩阵的一个例子，因而我们所选G的这组基对数值计算来说是非

常不好的. ■

6.8.4 标准正交系

内积空间中另一种处理逼近问题的方法是利用标准正交系. 内积空间中向量的有限序列或无穷序列 f_1, f_2, …, 如果满足

$$\langle f_i, f_j \rangle = 0 \qquad (i \neq j)$$

那么我们称 f_1, f_2, …是**正交的**. 如果对所有的 i 和 j,

$$\langle f_i, f_j \rangle = \delta_{ij}$$

那么这个集合被称为**标准正交的**. 由下列定理可知, 这样的正交系非常适合逼近问题.

定理 3(构造最佳逼近定理) 设$\{g_1, g_2, \cdots, g_n\}$是内积空间 E 中的一个标准正交系. 利用 $\sum_{i=1}^{n} c_i g_i$ 得到 f 的最佳逼近当且仅当 $c_i = \langle f, g_i \rangle$.

证明 设 G 是 $g_1, g_2, \cdots, g_n$ 生成的子空间. 正像定理 2 中刻画的那样, 最佳逼近 $\sum_{i=1}^{n} c_i g_i$ 等价于条件

$$f - \sum_{i=1}^{n} c_i g_i \perp G$$

因此, 我们只需要证明这个条件等价于条件 $c_i = \langle f, g_i \rangle$ 即可. 注意, $f - \sum_{i=1}^{n} c_i g_i$ 与 G 正交当且仅当它与每一个基向量 g_i 正交. 计算所需内积, 我们有$(1 \leqslant j \leqslant n)$

$$\begin{aligned}\langle f - \sum_{i=1}^{n} c_i g_i, g_j \rangle &= \langle f, g_j \rangle - \sum_{i=1}^{n} c_i \langle g_i, g_j \rangle \\ &= \langle f, g_j \rangle - c_j = 0\end{aligned}$$

■

下面是根据这些考虑所提出的方法: 如果希望用子空间 G 中的元去逼近 E 中的元, 首先要找到 G 的一组标准正交基$\{g_1, g_2, \cdots, g_n\}$, 那么 f 的最佳逼近就是 $\sum_{i=1}^{n} \langle f, g_i \rangle g_i$.

作为这一方法的说明, 我们重新考虑例 1

$$\sin x \approx c_1 x + c_2 x^3 + c_3 x^5$$

397

众所周知, 三维子空间的一组标准正交基可由以下三个勒让德多项式给出:

$$g_1(x) = x/\sqrt{2/3}$$

$$g_2(x) = (5x^3 - 3x)/(2\sqrt{2/7})$$

$$g_3(x) = (63x^5 - 70x^3 + 15x)/(8\sqrt{2/11})$$

因而我们问题的解是多项式 $\sum_{i=1}^{3} c_i g_i$, 其中 $c_i = \langle f, g_i \rangle$. 因此,

$$c_1 = \sqrt{3/2}\int_{-1}^{1} x \cdot \sin x \mathrm{d}x = 2\sqrt{3/2}(\alpha - \beta)$$

$$c_2 = \frac{1}{2}\sqrt{7/2}\int_{-1}^{1} \sin x(5x^3 - 3x)\mathrm{d}x = \sqrt{7/2}(-18\alpha + 28\beta)$$

$$c_3 = \frac{1}{8}\sqrt{11/2}\int_{-1}^{1} \sin x(63x^5 - 70x^3 + 15x)\mathrm{d}x = \frac{1}{4}\sqrt{11/2}(4\ 320\alpha - 6\ 728\beta)$$

其中 $\alpha=\sin 1$，以及 $\beta=\cos 1$. 用 α 和 β 的具体数值，我们求得近似解 $c_1\approx 0.738$，$c_2\approx -3.37\times 10^{-2}$，$c_3\approx 4.34\times 10^{-4}$. 因为已经存在可利用的标准正交基，所以用第二种方法很容易就得到这些结果. 其标准正交基是良态的(当然，也是最优的).

6.8.5 广义毕达哥拉斯法则和贝塞尔不等式

下面的结论是**广义毕达哥拉斯法则**.

引理 2(广义毕达哥拉斯法则) 若 $[g_1, g_2, \cdots, g_n]$ 是正交的，则

$$\left\| \sum_{i=1}^{n} a_i g_i \right\|^2 = \sum_{i=1}^{n} a_i^2 \| g_i \|^2$$

证明 $n=1$ 时结论显然成立. 如果等式对 n 成立，因为 $a_{n+1}g_{n+1}$ 与 $\sum_{i=1}^{n} a_i g_i$ 正交，应用基本的毕达哥拉斯法则：

$$\left\| \sum_{i=1}^{n+1} a_i g_i \right\|^2 = \left\| \sum_{i=1}^{n} a_i g_i \right\|^2 + \| a_{n+1} g_{n+1} \|^2$$

$$= \sum_{i=1}^{n} a_i^2 \| g_i \|^2 + a_{n+1}^2 \| g_{n+1} \|^2$$

因此，等式对 $n+1$ 也成立. ■

[398] 下列结论是**贝塞尔不等式**.

引理 3(贝塞尔不等式) 若 $[g_1, g_2, \cdots, g_n]$ 是标准正交的，则

$$\sum_{i=1}^{n} |\langle f, g_i \rangle|^2 \leqslant \| f \|^2$$

证明 设 $g^* = \sum_{i=1}^{n} \langle f, g_i \rangle g_i$. 由定理 3 知，$g^*$ 是由 g_i 生成的空间 G 中 f 的最佳逼近. 由定理 2 知，$f - g^* \perp G$. 根据毕达哥拉斯法则(用两次)，

$$\| f \|^2 = \| f - g^* \|^2 + \| g^* \|^2 \geqslant \| g^* \|^2 = \sum_{i=1}^{n} |\langle f, g_i \rangle|^2$$ ■

6.8.6 格拉姆-施密特过程

我们现在来关注如何获得标准正交基. 下面的定理描述了**格拉姆-施密特过程**. (回顾第 5 章，我们讨论过矩阵情况的格拉姆-施密特过程.)

定理 4(格拉姆-施密特过程定理) 设 $\{v_1, v_2, \cdots, v_n\}$ 是一个内积空间子空间 U 的基. 递归定义

$$u_i = \left\| v_i - \sum_{j=1}^{i-1} \langle v_i, u_j \rangle u_j \right\|^{-1} \left(v_i - \sum_{j=1}^{i-1} \langle v_i, u_j \rangle u_j \right) \qquad (i = 1, 2, \cdots, n)$$

则 $\{u_1, u_2, \cdots, u_n\}$ 是 U 的一个标准正交基.

证明 用 U_i 表示 $\{v_1, v_2, \cdots, v_i\}$ 生成的空间. 对 i 作数学归纳法，我们将证明：

1. $\{u_1, u_2, \cdots, u_i\} \subseteq U_i$.

2. $\{u_1, u_2, \cdots, u_i\}$ 标准正交.

对于 $i=1$，我们立即看出 $u_1 \in U_1$ 并且 $\| u_1 \| = 1$. 假设性质 1 和性质 2 对指标 $i-1$ 成立.

于是$\{u_1, u_2, \cdots, u_{i-1}\}\subseteq U_{i-1}$以及$v_i \notin U_{i-1}$. 因而$v_i$不是$u_1, u_2, \cdots, u_{i-1}$的线性组合，并且$u_i$定义中的分母不为0. 从而$u_i$有定义并且属于$U_i$. 因此，$\{u_1, u_2, \cdots, u_i\}\subseteq U_i$. 又因为$\sum_{j=1}^{i-1}\langle v_i, u_j\rangle u_j$是$U_{i-1}$中$v_i$的最佳逼近(根据定理3)，所以根据定理2，我们有$u_i \perp U_{i-1}$. 因此，有$u_i \perp \{u_1, u_2, \cdots, u_{i-1}\}$. 并且显然有$\|U_i\|=1$. ■

把格拉姆-施密特过程用于单项式函数1，x，x^2，…(按它们的自然顺序)时，会产生惊人的简化形式. 只要内积具有性质：对任意3个函数，有

399

$$\langle fg, h\rangle = \langle f, gh\rangle$$

那么下面的定理中就可以使用这种内积. 常用的内积

$$\langle f, g\rangle = \int_a^b f(x)g(x)w(x)\mathrm{d}x$$

显然符合上述条件.

定理5(正交多项式定理) 如下归纳定义的多项式序列是正交的：

$$p_n(x) = (x - a_n)p_{n-1}(x) - b_n p_{n-2}(x) \qquad (n \geqslant 2)$$

其中$p_0(x)=1$，$p_1(x)=x-a_1$，并且

$$a_n = \langle xp_{n-1}, p_{n-1}\rangle / \langle p_{n-1}, p_{n-1}\rangle$$

$$b_n = \langle xp_{n-1}, p_{n-2}\rangle / \langle p_{n-2}, p_{n-2}\rangle$$

证明 定理中的公式清楚地表明每个p_n都是首一n次多项式，从而都不是0. 因此，a_n和b_n公式中的分母都不为0. 下面我们对n作数学归纳法，证明$\langle p_n, p_i\rangle=0$，$1\leqslant i\leqslant n-1$. 而对$n=0$,则无需证明. 对$n=1$，由$a_1$的定义知

$$\langle p_1, p_0\rangle = \langle (x-a_1)p_0, p_0\rangle = \langle xp_0, p_0\rangle - a_1\langle p_0, p_0\rangle = 0$$

假设对指标$n-1$我们的断言正确，其中$n\geqslant 2$. 那么

$$\langle p_n, p_{n-1}\rangle = \langle xp_{n-1}, p_{n-1}\rangle - a_n\langle p_{n-1}, p_{n-1}\rangle - b_n\langle p_{n-2}, p_{n-1}\rangle = 0$$

$$\langle p_n, p_{n-2}\rangle = \langle xp_{n-1}, p_{n-2}\rangle - a_n\langle p_{n-1}, p_{n-2}\rangle - b_n\langle p_{n-2}, p_{n-2}\rangle = 0$$

对任一$i=0, 1, \cdots, n-3$，我们有

$$\begin{aligned}\langle p_n, p_i\rangle &= \langle xp_{n-1}, p_i\rangle - a_n\langle p_{n-1}, p_i\rangle - b_n\langle p_{n-2}, p_{n-i}\rangle \\ &= \langle p_{n-1}, xp_i\rangle \\ &= \langle p_{n-1}, p_{i+1} + a_{i+1}p_i + b_{i+1}p_{i-1}\rangle = 0\end{aligned}$$

在最后一步中，我们用了递归公式来表示xp_i. 如果$i=0$，应该记为$xp_0=p_1+a_1p_0$. ■

例2 利用定理5以及内积$\int_{-1}^{1} f(x)g(x)\mathrm{d}x$推导**勒让德多项式**.

解 最初几步计算如下进行：

$$p_0(x) = 1$$

$$a_1 = \langle xp_0, p_0\rangle / \langle p_0, p_0\rangle = 0$$

$$p_1(x) = x$$

$$a_2 = \langle xp_1, p_1\rangle / \langle p_1, p_1\rangle = 0$$

$$b_2 = \langle xp_1, p_0\rangle / \langle p_0, p_0\rangle = \frac{1}{3}$$

$$p_2(x) = x^2 - \frac{1}{3}$$

400

接下来的3个勒让德多项式是

$$p_3(x)=x^3-\frac{3}{5}x$$

$$p_4(x)=x^4-\frac{6}{7}x^2+\frac{3}{35}$$

$$p_5(x)=x^5-\frac{10}{9}x^3+\frac{5}{21}x$$

■

例 3 当使用下列内积时

$$\langle f,g\rangle=\int_{-1}^{1}f(x)g(x)\frac{\mathrm{d}x}{\sqrt{1-x^2}}$$

证明切比雪夫多项式构成一个[−1，1]上的正交系.

解 作变量替换 $x=\cos\theta$，把内积变为下列形式

$$\langle f,g\rangle=\int_{0}^{\pi}f(\cos\theta)g(\cos\theta)\mathrm{d}\theta$$

因为 $T_n(x)=\cos(n\cos^{-1}x)$，所以我们有(当 $n\neq m$)

$$\begin{aligned}\langle T_n,T_m\rangle&=\int_0^{\pi}\cos n\theta\cos m\theta\mathrm{d}\theta\\&=\frac{1}{2}\int_0^{\pi}[\cos(n+m)\theta+\cos(n-m)\theta]\mathrm{d}\theta\\&=\frac{1}{2}\left[\frac{\sin(n+m)\theta}{n+m}+\frac{\sin(n-m)\theta}{n-m}\right]_0^{\pi}=0\end{aligned}$$

■

6.8.7 算法

如果给定多项式 $u=\sum_{i=0}^{n}c_ip_i$，其中多项式 p_i 如定理5中所述，那么就可以如下高效地完成 $u(x)$ 的赋值：

```
d_{n+2} ← 0;  d_{n+1} ← 0
for k = n to 0 step −1 do
  d_k ← c_k + (x − a_{k+1}) d_{k+1} − b_{k+2} d_{k+2}
end do
```

401

从而，$u(x)=d_0$.

下面给出这个算法有效性的证明：

$$\begin{aligned}u(x)&=\sum_{k=0}^{n}c_kp_k(x)\\&=\sum_{k=0}^{n}[d_k-(x-a_{k+1})d_{k+1}+b_{k+2}d_{k+2}]p_k(x)\\&=d_0p_0(x)+d_1[p_1(x)-(x-a_1)p_0(x)]\\&\quad+\sum_{k=2}^{n}d_k[p_k(x)-(x-a_k)p_{k-1}(x)+b_kp_{k-2}(x)]\\&=d_0\end{aligned}$$

定理 6(极值性质定理) 定理 5 中所给出的多项式 p_n 具有性质：p_n 是首一 n 次多项式，并且因此 $\| p_n \|$ 取极小值.

证明 任意首一 n 次多项式可以表示成 $p_n - \sum_{i=0}^{n-1} c_i p_i$. 如果

$$p_n - \sum_{i=0}^{n-1} c_i p_i \perp \Pi_{n-1}$$

则这个函数的范数是最小值. 因为 $p_n \perp \Pi_{n-1}$，选取所有 $c_i = 0$，便可得到上述正交性关系. ■

如果$[u_1, u_2, \cdots, u_n]$是内积空间 E 中的正交系，那么下式定义了一族投影算子 P_n：

$$P_n f = \sum_{i=1}^{n} \langle f, u_i \rangle u_i$$

定理 7(正交投影定理) 算子 P_n 具有下列性质：

1. P_n 把 E 线性映射到 $u_1, u_2, \cdots, u_n$ 生成的子空间 U_n 上.
2. 每个 P_n 都是投影；即 $P_n^2 = P_n$.
3. 在 $f - P_n f \perp U_n$ 的意义下，每个 P_n 都是正交映射.
4. $P_n f$ 是 f 在 U_n 中的最佳逼近.
5. 每个 P_n 都是自伴的：$\langle P_n f, g \rangle = \langle f, P_n g \rangle$.

证明 证明作为习题 6.8.18 留给读者. ■ 402

6.8.8 格拉姆矩阵

内积空间中的逼近问题也可以用初等的方法求解而不使用标准正交基. 设$\{u_1, u_2, \cdots, u_n\}$是子空间 U 的任一组基，如果要计算 f 在 U 中的最佳逼近，我们可以这样做：为了使元 $u \in U$ 是 f 的最佳逼近，根据定理 2，其充分必要条件是 $u - f \perp U$. 而与之等价的一个条件是 $\langle u - f, u_i \rangle = 0$，$1 \leqslant i \leqslant n$. 令 $u = \sum_{j=1}^{n} c_j u_j$，我们得到条件

$$\sum_{j=1}^{n} c_j \langle u_j, u_i \rangle = \langle f, u_i \rangle \qquad (1 \leqslant i \leqslant n)$$

这些方程是这个问题的**正规方程**. 它是一个含有 n 个未知量 $c_1, c_2, \cdots, c_n$ 的 n 个线性方程的方程组. 其系数矩阵称为**格拉姆**矩阵，元素是 $G_{ij} = \langle u_i, u_j \rangle$.

引理 4(格拉姆矩阵引理) 若$\{u_1, u_2, \cdots, u_n\}$线性无关，则它的格拉姆矩阵非奇异.

证明 根据格拉姆-施密特过程，我们可以得到一个 $n \times n$ 矩阵 B 使得向量

$$v_i = \sum_{s=1}^{n} B_{is} u_s$$

构成一个标准正交集. 从而

$$\delta_{ij} = \langle v_i, v_j \rangle = \left\langle \sum_{s=1}^{n} B_{is} u_s, \sum_{r=1}^{n} B_{jr} u_r \right\rangle$$

这个等式的矩阵形式是 $I = BGB^{\mathrm{T}}$. 这就说明了 G 是非奇异的. ■

忍不住要给出这个结论的另外一种证明. 根据定理 2 后面的评注，上述逼近问题有唯一解 u. 由基的性质知，u 可由基唯一地表示成 $u = \sum c_i u_i$. 这说明我们的问题有唯一解$(c_1,$

$c_2, \cdots, c_n$). 因为这一点对任一 f 都是成立的，所以系数矩阵 G 一定非奇异.

引理4保证了正规方程有唯一的解. 因此，从理论上说子空间 U 的任一组基都可以用来求解逼近问题. 但在实际中，一定要考虑格拉姆矩阵的条件数. 回顾2.3节中关于希尔伯特矩阵的评注. 当所用内积是 $\langle f,g\rangle = \int_0^1 f(x)g(x)\mathrm{d}x$ 的时候，这个矩阵是函数 $u_j(x) = x^{j-1}$ 的格拉姆矩阵. 从条件的观点来看，最令人满意的是标准正交基，因为它们的格拉姆矩阵是单位矩阵.

403

习题 6.8

1. 用上确界范数求出区间 $[0, \pi/2]$ 上 $\sin x$ 的最佳逼近函数 $u(x)=\lambda x$. 提示：画出示意图. 函数 $\sin t-\lambda t$ 在区间 $(0, \pi/2)$ 内的点 ξ 上有极大值，并且它在 $\pi/2$ 上还有相同数值的极小值. 这些条件可以确定 ξ 和 λ.
2. (续)用一般平方范数求解上题.
3. 假设我们要用一个次数 $\leqslant n$ 的多项式去逼近一个偶函数，使用范数 $\| f\| = \left\{\int_{-1}^1 (f(x))^2\mathrm{d}x\right\}^{1/2}$. 证明：最佳逼近也是偶函数. 并推广上述结果.
4. 设 $p_0, p_1, p_2, \cdots$ 是一个多项式序列，使得(对每个 n)P_n 恰好是 n 次的. 证明：这个序列是线性无关的.
5. 证明帕塞瓦尔恒等式
$$\langle f,g\rangle = \sum_{i=1}^n \langle f,u_i\rangle\langle g,u_i\rangle$$
若 f 和 g 在标准正交集 $[u_1, u_2, \cdots, u_n]$ 生成的子空间中，则上式成立.
6. 著名的希尔伯特矩阵，其元素是
$$a_{ij} = (1+i+j)^{-1} \qquad (0\leqslant i,j\leqslant n)$$
证明：希尔伯特矩阵是函数 $1, x, x^2, \cdots, x^{n-1}$ 的格拉姆矩阵.
7. 在以 $\{v_1, v_2, \cdots, v_n\}$ 为基的向量空间中，其他任一组基都可以由下列线性变换得到
$$u_j = \sum_{i=1}^n a_{ij}v_i \qquad (1\leqslant j\leqslant n)$$
其中系数矩阵是非奇异的. 证明：在格拉姆–施密特过程中通过这种方式所形成的矩阵是上三角阵.
8. 在正交多项式的三项递归公式中，假设内积是 $\langle f,g\rangle = \int_{-a}^a f(x)g(x)w(x)\mathrm{d}x$，其中 w 是偶函数. 证明对所有的 n，$a_n = 0$. 并证明若 n 是偶数则 p_n 是偶函数；若 n 是奇数则 p_n 是奇函数.
9. 设 $\{v_1, v_2, \cdots, v_n\}$ 是内积空间中向量的正交集. 试问选择什么样的系数能产生 $\| f-\sum_{i=1}^n c_iv_i\|$ 的最小值？不要忽略 v 的某些系数也许为0的可能性.
10. 证明：在计算正交多项式线性组合的算法中，最多需要 $2n-1$ 次乘法运算. 而对于切比雪夫多项式的情况，有 n 次乘法运算就足够了.
11. 在正交多项式的三项递归公式中，证明：由 $b_n=\| p_{n-1}\|^2/\| p_{n-2}\|^2$ 给出的 b_n 是正的.
12. 证明：对于勒让德多项式，三项递归公式中的系数 $a_n=0$ 并且 $b_n=(n-1)^2/[(2n-1)(2n-3)]$.
13. 如果我们希望产生一个标准正交系，那么正交多项式的三项递归公式应该怎样变化？
14. 利用内积 $\langle u,v\rangle = \int_{-1}^1 u(x)v(x)\mathrm{d}x$，假设把格拉姆–施密特过程应用在函数序列 $x \mapsto (x^2-1)x^k (k=0,1,2,\cdots)$ 上. 证明：如果所得到的标准正交序列经过重新规范后构成首一多项式序列，那么这个序列满足下列形式的三项递归关系：

404

$$q_{n+1}(x) = xq_n(x) - b_nq_{n-1}(x)$$

给出 b_n 的一个公式，并求出前三个 q 多项式.

15. 证明：非零元的正交集必然线性无关.
16. 设 A 是内积空间上的线性变换，假设 A 是自伴的，也就是说 $\langle Af, g\rangle = \langle f, Ag\rangle$ 对所有的 f 和 g 都成立. 证明：对应于不同的 λ 值，方程 $Af=\lambda f$ 的解是相互正交的.
17. 设 $[u_1, u_2, \cdots]$ 是内积空间中的标准正交序列. 证明：对空间中任一 f，傅里叶系数 $\langle f, u_n\rangle$ 是平方可和的：

$$\sum_{n=1}^{\infty} \langle f, u_n\rangle^2 < \infty$$

18. 证明定理 7. 建议：重排那些性质的顺序为 1，4，3，2，5 会很有效.
19. 设 $\widetilde{T}_n$ 是 T_n 的首一多项式. 求出 $\widetilde{T}_0$，$\widetilde{T}_1$，…所满足的三项递归关系.
20. 求出 $\mathrm{dist}(f, G)$ 的公式，其中 G 是由标准正交集 $[g_1, g_2, \cdots, g_n]$ 生成的子空间.
21. 推导出下列勒让德多项式：

$$p_3(x) = x^3 - \frac{3}{5}x$$

$$p_4(x) = x^4 - \frac{6}{7}x^2 + \frac{3}{35}$$

$$p_5(x) = x^5 - \frac{10}{9}x^3 + \frac{5}{21}x$$

22. 直接利用定理 5，在 $[a, b]=[0, 1]$ 以及 $w(x)=1$ 的情况下，求出 p_0，p_1，p_2，p_3.
23.（续）借助于 p_3 与 Π_2 正交，确定形为 $p_3 = x^3 + Bx^2 + Cx + D$ 的 p_3. 并用上题验证你的结论.
24. 设计一个用计算机来计算 $\sum_{i=0}^{n} c_i p_i$ 的算法，按顺序计算出每个部分和 $\sum_{i=0}^{k} c_i p_i, k = 0,1,2,\cdots,n$. 假设定理 5 中的系数 a_k 和 b_k 是已知的.

6.9 最佳逼近：切比雪夫理论

在这一节，我们的讨论限制在空间 $C(X)$ 中，它是由定义在给定拓扑空间 X 上的全体实值连续函数所构成的空间. 我们假设 X 是一个紧豪斯多夫空间，希望避免考虑一般拓扑学的读者可以把 X 看作是实空间 $\mathbb{R}^k$ 中的一个有界闭集，例如，$\mathbb{R}$ 中的区间 $[a, b]$.

如果我们定义范数为

$$\| f \| = \max_{x \in X} | f(x) |$$

则空间 $C(X)$ 变成一个赋范空间（当然是一个巴拿赫空间），本节自始至终使用这种范数.

在空间 $C(X)$ 中有一类重要的最佳逼近问题如下：给定 $C(X)$ 的元 f，以及给定 $C(X)$ 中的一个有限维子空间 G，我们希望用 G 中的元尽可能好地去逼近 f. 因此（与上一节类似）定义

$$\mathrm{dist}(f, G) = \inf_{g \in G} \| f - g \|$$

405

并且询问是否存在最佳逼近；即是否存在元 $g \in G$ 使得

$$\| f - g \| = \mathrm{dist}(f, G)$$

这个问题已在 6.8 节（定理 1）得到了肯定的回答. 因此，在这样的背景下我们只考虑确定最佳逼近的问题. 为得到一些启发，先看一个例题.

例 1 描述用 Π_1 中的元在区间 $[0, \pi/2]$ 上最佳逼近函数 $f(x) = \cos x$.

解　Π_1 的元是线性函数，它的图形是直线．函数 $\cos x$ 以及离它不太远的一个典型的线性函数的图形如图 6-12 所示．

这个线性函数是否表示了一个最佳逼近？否，这个线性函数还可以下移一些，从而减少其极大偏差．此外，它的斜率还可以调整．对于最佳逼近，存在 3 个会出现极大偏差的点．显然它们是 0，$\pi/2$ 以及区间内一点 ξ．把极大偏差记为 δ，线性函数记为 $g(x)$，我们有

$$\begin{cases} g(0)-f(0)=\delta \\ g(\xi)-f(\xi)=-\delta \\ g(\pi/2)-f(\pi/2)=\delta \\ g'(\xi)-f'(\xi)=0 \end{cases}$$

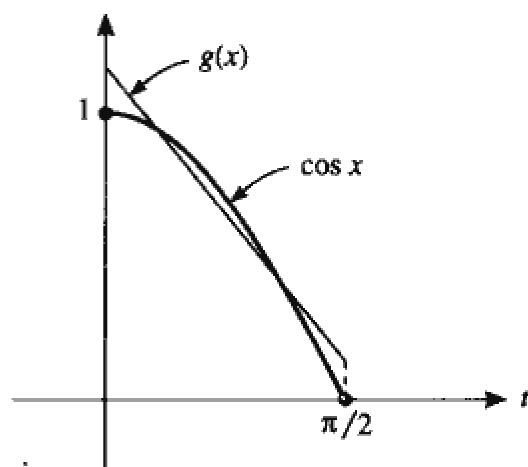


图 6-12　在区间$[0,\pi/2]$上逼近 $\cos x$

上面这 4 个方程用于确定 δ，ξ 以及等式 $g(x)=c_1+c_2x$ 中的两个系数．■

6.9.1　刻画最佳逼近的特征

在讨论最佳逼近的特征之前，先看一个结果

[406]

引理 1（最佳逼近性质引理）　对于 $g\in G$，下列性质等价：

1. g 是子空间 G 中 f 的最佳逼近．
2. 0 是子空间 G 中 $f-g$ 的最佳逼近．

证明　如果性质 1 正确，那么对所有 $h\in G$

$$\|f-g\|\leqslant\|f-g-h\|$$

这意味着 0 是 G 中 $f-g$ 的最佳逼近．反之，如果性质 2 正确，那么上述的不等式对所有 $h\in G$ 都成立，并且(因为 $g+h$ 可以是 G 中的任一元)性质 1 正确．■

因此，我们需要确切地理解：$C(X)$的元 f 以 G 中的 0 元作为最佳逼近；也就是具有性质 $\|f\|=\mathrm{dist}(f,G)$．对于 $f\in C(X)$，我们定义它的**临界集**是

$$\mathrm{crit}(f)=\{x\in X:|f(x)|=\|f\|\}$$

定理 1（科尔莫戈罗夫特征定理）　对 $C(X)$中的元 f 和子空间 G，下列性质等价：

1. $\|f\|=\mathrm{dist}(f,G)$．
2. G 中无元在 $\mathrm{crit}(f)$上与 f 有相同的符号．

证明　假设性质 1 错误．那么对某些 $g\in G$，$\|f-g\|<\|f\|$．对 $x\in\mathrm{crit}(f)$我们记 $\sigma(x)=\mathrm{sign}\,f(x)$并且有：

$$\sigma(x)[f(x)-g(x)]\leqslant|f(x)-g(x)|\leqslant\|f-g\|<\|f\|=|f(x)|=\sigma(x)f(x)$$

因此，$\sigma(x)g(x)>0$，并且 $g(x)$与 $f(x)$在 $\mathrm{crit}(f)$上有相同符号．

再假设性质 2 错误．设在 $\mathrm{crit}(f)$上 $g(x)f(x)>0$．不失一般性地，可以假设 $\|g\|=1$．因为 $\mathrm{crit}(f)$是紧集并且 gf 是连续函数，那么一定存在一个正数 ε 使得在 $\mathrm{crit}(f)$上 $g(x)f(x)>0$．记

$$\mathcal{O}=\{x\in X:g(x)f(x)>\varepsilon\}$$

因此 $\mathcal{O}$ 是一个包含 $\mathrm{crit}(f)$的开集．它的补集是与 $\mathrm{crit}(f)$不相交的紧集．从而，我们有

$$\rho \equiv \max\{|f(x)|: x \in X \setminus \mathcal{O}\} < \|f\|$$

现在我们试用λg去逼近f，其中系数λ要合理地选取. 让我们看看这需要什么条件. 首先，对$\mathcal{O}$中的点，我们将要求这个不等式按点成立

$$(f-\lambda g)^2 = f^2 - 2\lambda g f + \lambda^2 g^2 \leqslant \|f\|^2 - 2\lambda\varepsilon + \lambda^2 = \|f\|^2 - \lambda(2\varepsilon - \lambda) < \|f\|^2$$

容易证明，如果$0<\lambda<2\varepsilon$，该不等式成立. 对于剩余的点；即$X \setminus \mathcal{O}$中的点，我们要求

$$|f-\lambda g| \leqslant |f| + \lambda|g| \leqslant \rho + \lambda < \|f\|$$

如果$0<\lambda<\|f\|-\rho$，这个不等式成立. 因此，如果恰当地选取λ，就会有$\|f-\lambda g\|<\|f\|$，这说明性质1错误. ■ 407

根据引理1和科尔莫戈罗夫特征定理，我们得到下面的推论.

推论1(最佳逼近推论1，充分必要条件) 设f是$C(X)$中的元，G是$C(X)$的子空间，并且g^*是G中的元. g^*是G中f的最佳逼近的充分必要条件是不存在G中的元g在集合$\{x: |f(x)-g^*(x)| = \|f-g^*\|\}$上满足$g(x)[f(x)-g^*(x)]>0$.

推论2(最佳逼近推论2，充分必要条件) 元$g \in \Pi_1$是$f \in C[a, b]$的最佳逼近的充分必要条件是：函数$f-g$取值$\pm\|f-g\|$，同时要至少在$[a, b]$的3个点上交替变换符号.

证明 把科尔莫戈罗夫特征定理应用于$f-g$，则其特征性质是不存在Π_1中的元在$\text{crit}(f-g)$上与$f-g$有相同的符号. 那么$\text{crit}(f-g)$中一定至少存在3个点，在这些点上$f-g$的值交替变符号，因为否则，在$[a, b]$中有一点ξ，使得满足$f(x)-g(x)=\|f-g\|$的点在ξ的一边，而满足$f(x)-g(x)=-\|f-g\|$的点在ξ的另一边. 从而就存在一个在点ξ取零的线性函数，它在$\text{crit}(f-g)$上与$f-g$有相同的符号. ■

推论3(最佳逼近推论3，充分必要条件) 设X是$\mathbb{R}^2$中的有界闭集，G是由下列线性函数组成的$C(X)$的子空间：

$$g(x,y) = a + bx + cy$$

g是元$f \in G$的最佳逼近的充分必要条件是$f-g$的临界集一定包含图6-13所示的3种模式之一. 408

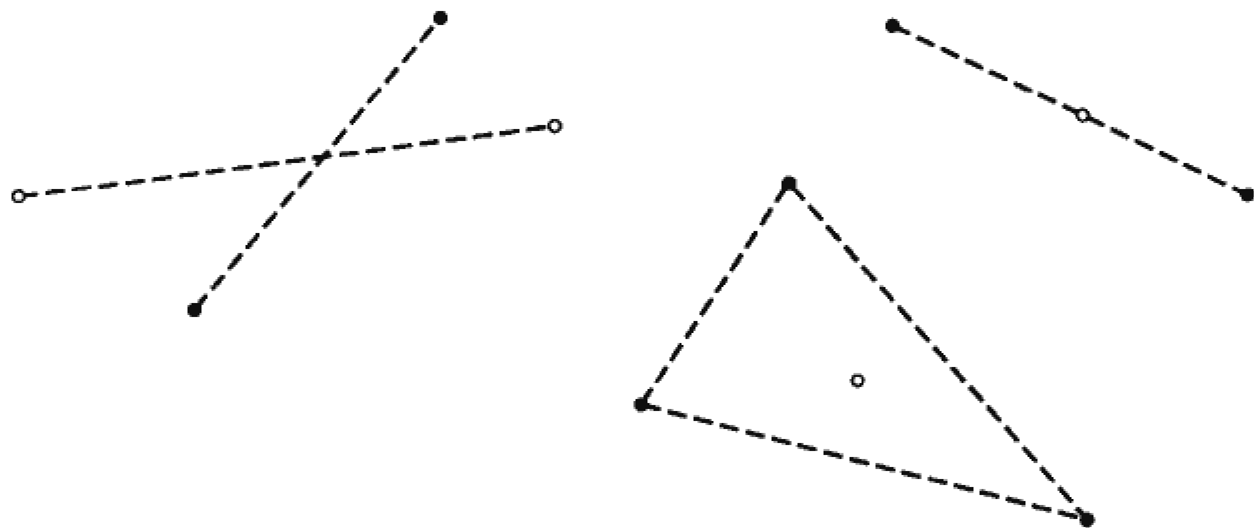


图6-13 最佳逼近中的临界点

6.9.2 凸性

线性空间的集合 K 称为**凸的**，如果它包含连接 K 中两点的每条线段. 严格地讲，它可以表述为

$$\left.\begin{array}{l} u,v\in K \\ 0\leqslant\theta\leqslant 1 \end{array}\right\}\Rightarrow \theta u+(1-\theta)v\in K$$

$\mathbb{R}^2$ 中的凸集和非凸集如图 6-14 所示.

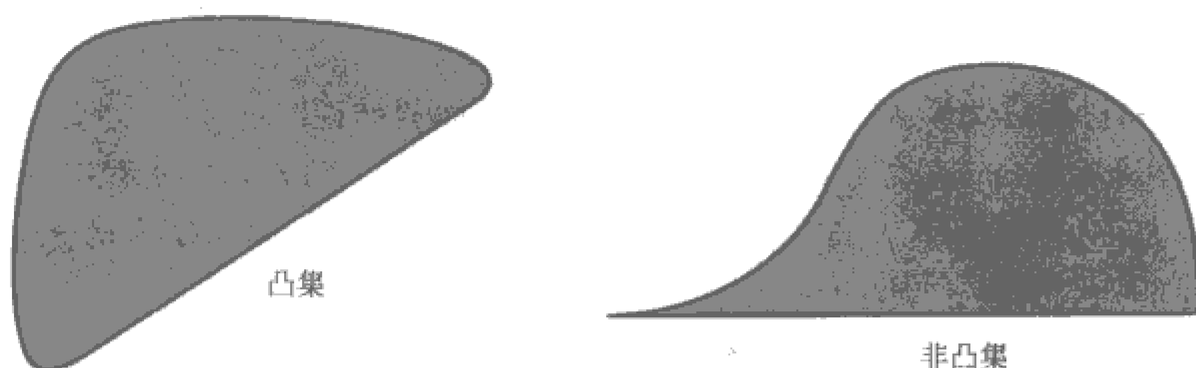


图 6-14 凸集和非凸集

当 $\sum_{i=1}^{k}\theta_i=1$ 且 $\theta_i\geqslant 0$ 时，形如 $\sum_{i=1}^{k}\theta_i u_i$ 的向量的线性组合称为**凸组合**. 从给定的集合 S 中选取点的所有凸组合的集合称为 S 的**凸包**. 从而，我们有

$$\mathrm{co}(S)=\left\{\sum_{i=1}^{k}\theta_i u_i : k\in\mathbb{N},u_i\in S,\theta_i\geqslant 0,\sum_{i=1}^{k}\theta_i=1\right\}$$

引理 2（闭凸集性质引理） 设 K 是 $\mathbb{R}^n$（或任一希尔伯特空间）中的闭凸集. 则 K 包含唯一的极小范数点. 进而，K 的下列性质等价：

1. $0\notin K$.

2. 存在一个向量 v 使得对所有 $u\in K$，都有 $\langle v,u\rangle>0$.

证明 设 $\rho=\inf\{\|u\| : u\in K\}$，选取 $u_j\in K$ 使得 $\|u_j\|\to\rho$. 由平行四边形法则给出：

$$\begin{aligned}\|u_i-u_j\|^2&=2\|u_i\|^2+2\|u_j\|^2-\|u_i+u_j\|^2\\&=2\|u_i\|^2+2\|u_j\|^2-4\|(u_i+u_j)/2\|^2\\&\leqslant 2\|u_i\|^2+2\|u_j\|^2-4\rho^2\to 2\rho^2+2\rho^2-4\rho^2=0\end{aligned}$$

所以序列$[u_j]$是柯西序列. 由空间的完备性知柯西序列收敛，记为 $u_j\to u$. 因为 K 是闭的，所以 $u\in K$. 再由连续性知 $\|u\|=\rho$. 唯一性的证明可利用类似于上面的理由：如果 u 和 u' 是 K 中范数都为 ρ 的两个元，那么

409

$$\|u-u'\|^2\leqslant 2\|u\|^2+2\|u'\|^2-4\rho^2=0$$

如果性质 2 成立，显然立即可得性质 1. 反之，假设性质 1 正确，令 v 是 K 中最小范数点. 对任意 $u\in K$ 和 $\theta\in(0,1)$，我们有

$$\begin{aligned}0\leqslant\|\theta u+(1-\theta)v\|^2-\|v\|^2&=\|\theta(u-v)+v\|^2-\|v\|^2\\&=\theta^2\|u-v\|^2+2\theta\langle u-v,v\rangle\end{aligned}$$

这就表明

$$0 \leqslant \theta \| u - v \|^2 + 2\langle u - v, v\rangle$$

令 $\theta \downarrow 0$，我们得到 $\langle u-v, v\rangle \geqslant 0$，由此 $\langle v, u\rangle \geqslant \langle v, v\rangle > 0$. ■

定理 2(卡拉泰奥多里定理) 设 S 是 n 维线性空间的子集. 则 S 凸包中的每一个点都是 S 中至多 $n+1$ 个点的凸组合.

证明 设 p 是 S 凸包中的一点. 不失一般性，取 p 是 0. 则对适当的 $\theta_i \in [0,1]$，$u_i \in S$，以及 $\sum_{i=1}^{k}\theta_i = 1$，我们有 $0 = \sum_{i=1}^{k}\theta_i u_i$. 假设选取了 0 的最小的这种表达式(即 k 尽可能的小). 那么必然有 $\theta_i > 0, 0 \leqslant i \leqslant k$. 如果 $k \leqslant n+1$，则证明结束. 如果 $k > n+1$，那么找一个非平凡线性相关组合 $\sum_{i=2}^{k}\lambda_i u_i = 0$. 取 $\lambda_1 = 0$，可以看出对所有的实数 t，有 $\sum_{i=1}^{k}(\theta_i + t\lambda_1)u_i = 0$. 设

$$\phi(t) = \min_{1\leqslant i\leqslant k}(\theta_i + t\lambda_i)$$

函数 ϕ 连续且 $\phi(0) > 0$. 对于某些 t，$\phi(t) < 0$. 因此，存在一个特殊的 t_0 使得 $\phi(t_0) = 0$. 令 $\theta'_i = \theta_i + t_0\lambda_i$，我们有 $\theta'_i \geqslant 0$，$\min \theta'_i = 0$，并且 $\theta'_1 = \theta_1 > 0$. 从而在表达式 $\sum_{i=1}^{k}\theta'_i u_i = 0$ 中，有一项 $\theta'_i = 0$. 将等式两端除以 $\sum\theta'_i$，我们就把 0 表示成了 S 中 $k-1$ 个元的凸组合，而这与 k 的最小性假设矛盾. ■

引理 3(紧集的凸包引理) 有限维赋范线性空间中紧集的凸包是紧的.

证明 设 S 是 n 维空间 X 的紧集. 考虑所有 $(2n+2)$ 维元的集合 V：

$$v \equiv (\theta_0, \theta_1, \cdots, \theta_n, u_0, u_1, \cdots, u_n)$$

其中 $\theta_i \in \mathbb{R}$，$u_i \in S$，$\theta_i \geqslant 0$ $\sum_{i=0}^{n}\theta_i = 1$. 这些 $(2n+2)$ 维元是 $\mathbb{R}^{n+1}\times X \times X \times \cdots \times X$ 中的点. 这个空间维数是 $n+1+(n+1)n = (n+1)^2$，并且集合 V 是有界闭集；因此，V 是紧的. 定义 $f(v) = \sum_{i=0}^{n}\theta_i u_i$. 则 f 是 V 到 $\mathrm{co}(S)$ 的映射. 根据卡拉泰奥多里定理，f 是满射(到上的). 同时它也是连续的. 因为紧集的连续映像是紧的，所以 $\mathrm{co}(S)$ 是紧的. ■ [410]

定理 3(线性不等式定理) 对于 $\mathbb{R}^n$ 中的紧集 S，下列性质等价：

1. 存在一点 v 使得对所有的 $u \in S$，都有 $\langle v, u\rangle > 0$.

2. 0 不在 S 的凸包中.

证明 如果性质 2 不成立，那么对适当的 $\theta_i > 0$ 和 $u_i \in S$，我们有表达式 $0 = \sum_{i=1}^{k}\theta_i u_i$. 对任意的 $v \in \mathbb{R}^n$，我们有

$$0 = \langle v, 0\rangle = \sum_{i=1}^{k}\theta_i\langle v, u_i\rangle$$

显然，不可能所有的 $\langle v, u_i\rangle$ 都是正的. 因此，性质 1 不成立.

反之，假设性质 2 正确. 由引理 3 知，$\mathrm{co}(S)$ 是紧的. 从而可由引理 2 知. 性质 1 正确. ■

6.9.3 线性方程组的切比雪夫解

推论 4(最佳逼近推论 4，充分必要条件) 设 A 是一个 $m\times n$ 矩阵，$b\in\mathbb{R}^m$，$x\in\mathbb{R}^n$，$\sigma_i=\text{sign}(Ax-b)_i$，且 $I=\{i:|(Ax-b)_i|=\|Ax-b\|_\infty\}$. x 极小化范数 $\|Ax-b\|_\infty$ 的充分必要条件是 0 在集合 $\{\sigma_i A_i: i\in I\}$ 的凸包中，其中 A_i 表示 A 的第 i 行.

证明 为了极小化 $\|Ax-b\|_\infty$，我们要寻找 A 的列向量的一个线性组合，使得它尽可能地靠近 b. 我们把所有列向量看作集合 $T=\{1, 2, \cdots, n\}$ 上的函数. 被逼近的元是 b，逼近的子空间 G 是 A 的列空间. 根据科尔莫戈罗夫特征定理，解 x 具有特征：不存在 G 中的元在 I 上的符号是 σ_i. 等价的说法是：系统 $\sigma_i(Av)_i>0$，$i\in I$ 是矛盾的. 因为该系统可表述成 $\langle\sigma_i A_i, v\rangle>0$，$i\in I$. 因此定理 3 给出了一个与之等价的条件；即 $0\in\text{co}\{\sigma_i A_i: i\in I\}$. ■

例 2 利用推论 4，确定 $x=(2, 3)^{\mathrm{T}}$ 是否为下列方程组的切比雪夫解.

$$\begin{cases} 5x_1-7x_2=14 \\ 3x_1+x_2=8 \\ x_1-9x_2=-23 \\ 3x_1-1x_2=6 \\ 6x_1-x_2=6 \end{cases}$$

解 残差 $r_i=\langle A_i, x\rangle-b_i$ 是 3，1，−2，−3，3. 从而，σ_i 是 +1，+1，−1，−1，+1. 临界集是 $I=\{1, 4, 5\}$. 根据推论 4，x 是切比雪夫解当且仅当

411

$$0\in\text{co}\{\sigma_1 A_1, \sigma_4 A_4, \sigma_5 A_5\}$$

可以很快地确定这三个向量的坐标，并表明这时结论成立. 如图 6-15 所示.

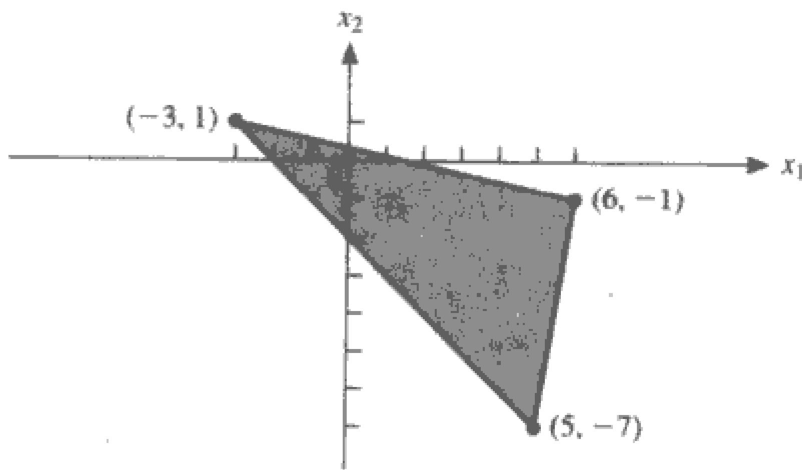


图 6-15 点 0 在三点的凸包中

■

6.9.4 再论特征定理

在下面的定理中，我们要用到一些特殊的符号. 如果 G 是 $C(X)$ 的 n 维子空间，且 $\{g_1, g_2, \cdots, g_n\}$ 是 G 的一组基，我们记

$$\vec{g}(x)=(g_1(x), g_2(x), \cdots, g_n(x)) \qquad (x\in X)$$

这可以看作 $\mathbb{R}^n$ 中的点.

对任一 $x\in X$，我们用下列等式定义 $C(X)$ 上的线性泛函 $\hat{x}$

$$\hat{x}(f)=f(x)\qquad (f\in C(X))$$

并通过简单的计算可得到 $\hat{x}$ 的线性性：

$$\hat{x}(\alpha f+\beta g)=(\alpha f+\beta g)(x)=\alpha f(x)+\beta g(x)=\alpha\hat{x}(f)+\beta\hat{x}(g)$$

定理 4(特征定理) 设 G 是 $C(X)$ 的 n 维子空间，且 f 是 $C(X)$ 的元，则下列性质等价：

1. $\|f\|=\mathrm{dist}(f,G)$.
2. 不存在 G 中的元在 $\mathrm{crit}(f)$ 上与 $f(x)$ 有相同的符号.
3. 0 在集合 $\{f(x)\vec{g}(x):x\in \mathrm{crit}(f)\}$ 的凸包中.
4. 存在一个泛函 $\sum_{i=1}^{k}\lambda_i\hat{x}_i$，它零化 G 并且满足条件：$x_i\in\mathrm{crit}(f),\lambda_i f(x_i)>0,k\leqslant n+1$.

证明 性质 1⇒性质 2 是前面已证明的科尔莫戈罗夫特征定理的一部分.

性质 2⇒性质 3，假设性质 2 成立. 设 $\{g_1, g_2, \cdots, g_n\}$ 是 G 的任一基. 那么在 $\mathrm{crit}(f)$ 上，我们不可能找到 $c_1, c_2, \cdots, c_n$ 使得

$$\sum_{i=1}^{n}c_i f(x)g_i(x)>0$$

因为这个不等式可写成 $\langle c, f(x)\vec{g}(x)\rangle>0$，所以由线性不等式定理知，0 在下列点集的凸包中： [412]

$$\{f(x)\vec{g}(x):x\in\mathrm{crit}(f)\}$$

性质 3⇒性质 4，假设性质 3 成立. 根据卡拉泰奥多里定理，我们有

$$0=\sum_{i=1}^{k}\theta_i f(x_i)\vec{g}(x_i)$$

其中 $k\leqslant n+1,\theta_i>0,x_i\in\mathrm{crit}(f)$. 令 $\lambda_i=\theta_i f(x_i)$，我们有 $\lambda_i f(x_i)>0$ 并且 $0=\sum_{i=1}^{k}\lambda_i\vec{g}(x_i)$. 从 $\vec{g}$ 的定义知，最后这个等式表明：对于 $j=1,2,\cdots,n,0=\sum_{i=1}^{k}\lambda_i g_j(x_i)=\left(\sum_{i=1}^{k}\lambda_i\hat{x}_i\right)(g_j)$. 因为函数 g_j 生成 G，所以我们看出泛函 $\sum_{i=1}^{k}\lambda_i\hat{x}_i$ 零化 G.

性质 4⇒性质 1，假设性质 4 成立. 如果 $h\in G$，那么

$$\|f\|\sum_{i=1}^{k}|\lambda_i|=\sum_{i=1}^{k}\lambda_i f(x_i)=\sum_{i=1}^{k}\lambda_i[f(x_i)-h(x_i)]\leqslant\|f-h\|\sum_{i=1}^{k}|\lambda_i|$$

因此，$\|f\|\leqslant\|f-h\|$，从而性质 1 成立. ■

例 3 设 $X=[0, 1]$ 并且 G 由 $g_1(x)=1-2x^2$ 和 $g_2(x)=x-x^2$ 生成. 证明：任一 $f\in C[0, 1]$，如果 $\|f\|=f(0)=f(1)$，那么一定有 $\|f\|=\mathrm{dist}(f, G)$.

解 利用定理 4 的性质 4，对于 $g\in G$，我们有 $g(0)+g(1)=0$. (简单检验即可知两个基函数具有此性质.) 因此，我们可取定理中的 $x_1=0$，$x_2=1$，$\lambda_1=1$，$\lambda_2=1$. ■

6.9.5 哈尔子空间

定义 1(哈尔子空间定义) 设 G 是 $C(X)$ 中的 n 维子空间. 如果 G 的非零元在 X 中没有 n 个或者更多的零点，那么 G 称为**哈尔子空间**.

例 4 设 X 是 $\mathbb{R}$ 的子集合，则次数 $<n$ 的多项式的子空间 Π_{n-1} 是 $C(X)$ 中的一个哈尔子空间. ■

下面的结果表明哈尔子空间非常适合插值问题的要求.

引理 4(引理) $C(X)$ 中的 n 维子空间 G 是哈尔子空间的充要条件是对于任意的实数 λ_1, λ_2, $\cdots$, λ_n 和 X 中任意的互异点 x_1, x_2, $\cdots$, x_n, 存在 G 中的唯一元 g 使得 $g(x_i)=\lambda_i(1\leqslant i\leqslant n)$.

[413] **证明** 选取 G 的一组基 $\{g_1, g_2, \cdots, g_n\}$. 那么插值问题就是确定 c_1, c_2, $\cdots$, c_n 使得

$$\sum_{j=1}^{n} c_j g_j(x_i) = \lambda_i \qquad (1 \leqslant i \leqslant n)$$

如果对任意选取的 λ_i 这个问题都有解，那么以 $g_j(x_i)$ 为元素的矩阵非奇异，并且所对应的齐次问题(所有 $\lambda_i=0$)只有 0 解(对全部的 i, $c_i=0$). 因此，不存在 g_1, g_2, $\cdots$, g_n 的线性组合在点 x_1, x_2, $\cdots$, x_n 处为零. 因为点 x_1, x_2, $\cdots$, x_n 是任意的，所以 G 的哈尔性质成立. 由于上述讨论过程是可逆的，故证明完毕. ■

对哈尔子空间，以下列方式描述最佳逼近：

定理 5(哈尔子空间中最佳逼近定理) 设 G 是 $C(X)$ 中 n 维哈尔子空间，且 $f \in C(X)$. 则 $\| f \| = \text{dist}(f,G)$ 的充分必要条件是存在形如 $\sum_{i=1}^{n+1}\lambda_i\hat{x}_i$ 的泛函零化 G 并且满足 $x_i \in \text{crit}(f)$ 和 $\lambda_i f(x_i) > 0$.

证明 在用 k 替换 $n+1$ 的情况下，定理 4 给出了同样的充分必要条件. 当 $k \geqslant n+1$ 时，等式 $\sum_{i=1}^{k}\lambda_i g(x_i) = 0$ 只能平凡地成立. 我们来证明这一点. 设 $k \leqslant n$，用 G 的插值性质(引理 4) 来选择 $g \in G$ 使得 $g(x_i) = \lambda_i$. 因而，我们有 $0 = \sum_{i=1}^{k}\lambda_i g(x_i) = \sum_{i=1}^{k}\lambda_i^2$. ■

6.9.6 最佳逼近的唯一性

定理 6(强唯一性定理) 设 G 是 $C(X)$ 中有限维哈尔子空间，并且 $C(X)$ 的元 f 使得 $\| f \| = \text{dist}(f, G)$，则存在一个正常数 γ(与 f 有关)使得对所有 $g \in G$，都有 $\| f+g \| \geqslant \| f \| + \gamma \| g \|$.

证明 设 G 的维数是 n. 根据定理 5，存在 $\text{crit}(f)$ 中的点 $x_1, x_2, \cdots, x_n$ 以及正系数 θ_i 使得对所有 $g \in G$，都有 $\sum_{i=0}^{n}\theta_i\sigma_i g(x_i) = 0$，其中 $\sigma_i = \text{sign} f(x_i)$. 设 h 是 G 中范数是 1 的元. 因为 $\sum_{i=0}^{n}\theta_i\sigma_i h(x_i) = 0$，所以至少有一个 $\sigma_i h(x_i)$ 是正数. 因此 $\max_i \sigma_i h(x_i) > 0$. 因为这个表达式连续并且 G 的单位胞腔曲面是紧的，所以推得

$$\gamma \equiv \inf_h \max_i \sigma_i h(x_i) > 0$$

如果 $g \in G$ 并且 $g \neq 0$，那么对某些 i，$\sigma_i g(x_i)/\| g \| \geqslant \gamma$. 因此，我们有

[414]

$$\| f+g \| \geqslant \sigma_i f(x_i) + \sigma_i g(x_i) \geqslant \| f \| + \gamma \| g \|$$ ■

推论 5(推论) 若 G 是 $C(X)$ 中有限维哈尔子空间，则 $C(X)$ 的每个元在 G 中有唯一的最佳逼近.

证明 设 g 是 $C(X)$ 的元 f 的最佳逼近，那么 $f-g$ 在 G 中有 0 作为最佳逼近. 如果 $h\in G$ 并且 $h\neq 0$，根据强唯一性定理，有

$$\|f-g+h\|\geqslant\|f-g\|+\gamma\|h\|>\|f-g\| \qquad \blacksquare$$

定理 7(连续性定理) 设 G 是 $C(X)$ 中有限维哈尔子空间. 设映射 $A: C(X)\to G$，使得 $\|f-Af\|=\mathrm{dist}(f, G)$，则对每个 f 存在一个正数 $\lambda(f)$ 使得

$$\|Af-Ah\|\leqslant\lambda(f)\|f-h\| \qquad h\in C(X)$$

证明 根据强唯一性定理 6，存在一个正数 $\gamma(f)$ 使得对所有 $g\in G$，

$$\|f-g\|\geqslant\|f-Af\|+\gamma(f)\|Af-g\|$$

令 $g=Ah$，我们有

$$\begin{aligned}\gamma(f)\|Af-Ah\| &\leqslant \|f-Ah\|-\|f-Af\|\\ &\leqslant\|f-h\|+\|h-Ah\|-\|f-Af\|\\ &\leqslant\|f-h\|+\|h-Af\|-\|f-Af\|\\ &\leqslant\|f-h\|+\|h-f\|+\|f-Af\|-\|f-Af\|\\ &=2\|f-h\|\end{aligned} \qquad \blacksquare$$

6.9.7 切比雪夫交替定理

引理 5(引理) 设 G 是 $C[a,b]$ 中 n 维哈尔子空间. 假设有 $n+1$ 个点使得 $a\leqslant x_0<x_1<\cdots<x_n\leqslant b$ 并且对每个 $g\in G$ 都有 $\sum_{i=0}^{n}\lambda_i g(x_i)=0$. 若 $\sum_{i=0}^{n}|\lambda_i|\neq 0$，则 λ_i 交替变符号：$\lambda_i\lambda_{i-1}<0, i=1, 2,\cdots,n$.

证明 对任一 $j\in\{1, 2, \cdots, n\}$ 我们实施下述论证过程：根据引理 2，G 中存在唯一元 g_j 具有插值性质：

$$\begin{cases} g_j(x_i)=0 & \text{对 } 0\leqslant i\leqslant j-2 \text{ 及 } j+1\leqslant i\leqslant n\\ g_j(x_j)=1 \end{cases}$$

415

因而我们有

$$0=\sum_{i=0}^{n}\lambda_i g_j(x_i)=\lambda_{j-1}g_j(x_{j-1})+\lambda_j \tag{1}$$

如果 $g_j(x_{j-1})=0$ 或者 $g_j(x_{j-1})<0$，可知 g_j 有 n 个零点，这与哈尔子空间性质矛盾. 所以 $g_j(x_{j-1})>0$. 等式(1)表明：如果任一 λ_j 是 0，那么所有的 λ_i 都是 0，这与假设条件矛盾. 因而，由等式(1)知 λ_j 和 λ_{j-1} 符号相反. $\blacksquare$

定理 8(切比雪夫交替定理) 设 G 是 $C[a, b]$ 中 n 维哈尔子空间，且 X 是 $[a, b]$ 中的一个闭集. 对元 $f\in C(X)$，下列性质等价：

1. $\|f\|=\mathrm{dist}(f, G)$；即，$f$ 在 G 中有 0 作为最佳逼近.
2. 存在 X 中的点 $x_0<x_1<\cdots<x_n$ 使得 $f(x_{i-1})f(x_i)=-\|f\|^2$，$1\leqslant i\leqslant n$.

证明 根据定理 7，性质 1 等价于：

3. 存在 $\mathrm{crit}(f)$ 中的点 $x_0, x_1, \cdots, x_n$，以及系数 $\lambda_0, \lambda_1, \cdots, \lambda_n$ 使得 $x_0<x_1<\cdots<x_n$，对所有 $g\in G$ 都有 $\sum_{i=0}^{n}\lambda_i g(x_i)=0$，以及 $\lambda_i f(x_i)>0$.

如果性质 3 成立，由引理 5 知系数 λ_i 一定交替变号. 因为 $\lambda_i f(x_i)>0$，所以 $f(x_i)$ 也交替变号，因此，性质 2 成立.

反之，如果性质 2 成立，显然每个 x_i 都是 f 的临界点，并且 $f(x_i)$ 交替变号. 假设确定 sign $f(x_i)=(-1)^i$. 如果性质 1 不成立，那么对某些 $g\in G$，$\|f-g\|<\|f\|$. 因此，我们有

$$(-1)^i[f(x_i)-g(x_i)]\leqslant\|f-g\|<\|f\|=(-1)^i f(x_i)$$

从而 $(-1)^i g(x_i)>0$，并且 g 至少有 n 个零点，这与 G 的哈尔性质发生矛盾. 注意，g 的这些零点一定会在 $[a, b]$ 中，但不一定在 X 中，这就是在 $[a, b]$ 上需要哈尔条件的原因. ■

推论 6（切比雪夫交替推论） 设 G 是 $C[a, b]$ 中 n 维哈尔子空间，并设 $f\in C[a, b]$ 以及 $g\in G$，则 g 是 f 在 G 中最佳逼近的充分必要条件是存在 $[a, b]$ 中的点 $x_0<x_1<\cdots<x_n$ 使得

$$f(x_i)-g(x_i)=(-1)^i c\|f-g\| \qquad (0\leqslant i\leqslant n,\quad |c|=1)$$

6.9.8 算法

[416] 在本节的剩余部分，我们将讨论切比雪夫逼近的计算问题并且概要介绍一些有效算法. 我们把下述表达式的极小化作为基本问题：

$$\Delta(c)=\left\|f-\sum_{i=1}^{n}c_i g_i\right\|_\infty=\sup_{x\in X}\left|f(x)-\sum_{i=1}^{n}c_i g_i(x)\right|$$

在这个问题中 f，g_1，g_2，…，g_n 是 $C(X)$ 中给定的函数，我们希望确定系数向量 $c=(c_1, c_2, \cdots c_n)$ 使得 $\Delta(c)$ 尽可能小.

这个问题的一个基本算法是迭代算法，在其每一步迭代中它都需要求解一个更简单的同类问题. 这个算法称为**列梅兹第一算法**；我们现在来描述这个算法.

用 G 表示由 $\{g_1, g_2, \cdots, g_n\}$ 生成的子空间. 在算法的第 k 步，给出 X 中的一个有限子集 X_k. 通过等式

$$\|f\|_k=\max_{x\in X_k}|f(x)|$$

用这个有限集来定义 $C(X)$ 中的一个拟范数，在这步迭代中，需要确定（所用方法将在后面讨论）G 中的元 h_k 极小化拟范数

$$\|f-h_k\|_k$$

这是另一个切比雪夫逼近问题，但是它仅包含有限点集 X_k. 下一步，在 X 中选取一点 x_k 使得

$$|f(x_k)-h_k(x_k)|=\|f-h_k\|$$

把这一点添加在 X_k 中就构成了下一集合 X_{k+1}，并且重复这一过程. 在第 k 步所得到的向量 h_k 通常是下一步寻找 h_{k+1} 的一个好起点.

定理 9（列梅兹算法定理） 若在列梅兹第一算法中的初始拟范数 $\|\ \|_1$ 是 G 上的真范数，则 $\lim\limits_{k\to\infty}\|f-h_k\|=\operatorname{dist}(f, G)$. 序列 $[h_k]$ 有聚点，并且每个聚点都是 f 的一个最佳逼近.

证明 直接从范数的定义和包含关系 $X_1\subseteq X_2\subseteq\cdots$ 知，对于 $1\leqslant k\leqslant i$ 以及 $g\in G$，我们有

$$\|f-g\|_1\leqslant\|f-g\|_k\leqslant\|f-g\|_i\leqslant\|f-g\|$$

由此不等式可推出

$$\|f-h_k\|_1\leqslant\|f-h_k\|_k\leqslant\|f-h_i\|_i\leqslant\operatorname{dist}(f,G)$$

这表明序列 $\|h_k\|_1$ 有界. 因为有限维空间 G 上的所有范数等价. 所以序列 $\|h_k\|$ 有界.

因此，h_k 的某些子序列收敛于点 h^*. 给定 $\epsilon>0$，选取 k 使得 $\| h_k-h^* \|<\epsilon$. 再选取 $i>k$ 使得 $\| h_i-h^* \|<\epsilon$. 于是我们有

$$\begin{aligned}\operatorname{dist}(f,G) &\leqslant \| f-h^* \| \leqslant \| f-h_k \| + \| h_k-h^* \| \\ &\leqslant | f(x_k)-h_k(x_k) | + \epsilon \\ &\leqslant \| f-h_k \|_i + \epsilon \\ &\leqslant \| f-h_i \|_i + \| h_i-h^* \|_i + \| h^*-h_k \|_i + \epsilon \\ &\leqslant \operatorname{dist}(f,G) + 3\epsilon\end{aligned}$$

因为 ϵ 任意，所以我们断定 $\| f-h^* \| = \operatorname{dist}(f, G)$. 这样，序列[$h_k$]的任一聚点都是 G 中 f 的一个最佳逼近.

接下来要证明序列 $d_k=\| f-h_k \|$ 收敛于 $\operatorname{dist}(f, G)$. 由于这个序列有界，因此它有收敛的子序列. 设 $d_{k_i}\to d^*$. 再设 h'表示子序列[h_{k_i}]的聚点. 根据证明的第一部分，$d^*=\| f-h' \|=\operatorname{dist}(f, G)$. 这说明序列[$d_k$]只有一个聚点；那就是 $\operatorname{dist}(f, G)$. 所以，我们有 $d_k\to\operatorname{dist}(f,G)$. ■

关注列梅兹第一算法的每一步都是有用的，很容易计算未知数 $d^*=\operatorname{dist}(f, G)$的上界和下界. 事实上，我们有

$$\| f-h_k \|_k \leqslant d^* \leqslant \min_{1\leqslant i\leqslant k} \| f-h_i \|$$

这个下界单调增加收敛于 d^*，而其上界单调减少收敛于 d^*. 最简单的上界 $\| f-h_k \|$ 也收敛于 d^*，但它并不总是单调的.

下面我们讨论这算法的一个变形，称为**交换方法**. 为遏制集合 X_k 增长，算法中每当添加一个新元的同时删除一个旧元. 在实际问题中，通常采用这个方法，尽管它的有效性还依赖于有关基函数 g_1, g_2, …, g_n 的更多假设条件. 删除一个元和添加一个元的过程称为一个**交换**. 如果初始子集 X_1 包含 $n+1$ 个元，并且算法的每一步都发生一个交换，那么每个 X_k 恰好包含 $n+1$ 个元.

根据最佳逼近的特征定理，我们知道在算法的第 k 步，$\mathbb{R}^n$的原点将落在这 $n+1$ 个向量的凸包中

$$e_k(x)(g_1(x),g_2(x),\cdots,g_n(x)) \qquad x\in X_k \tag{2}$$

其中 $e_k(x)=f(x)-h_k(x)$. 有了 h_k，就可以选取前面称为 x_k 的点，即 x_k 是函数 $f-h_k$ 的临界点. 然后用 x_k 替换 X_k 中的一个点，这个交换用这样一种方式完成，它要求$\mathbb{R}^n$的原点保留在(2)式中所提及的向量凸包中. 从而向量

$$e_k(x_k)(g_1(x_k),g_2(x_k),\cdots,g_n(x_k))$$

将替换前面 $n+1$ 个向量集合中的一个. 下面给出控制交换及展示如何实现这种交换的定理.

417
～
418

定理 10(交换定理)　设$\{u_0, u_1, \cdots, u_{n+1}\}$是$\mathbb{R}^n$中 $n+2$ 个点的集合. 若 0 位于$\{u_0, u_1, \cdots, u_{n+1}\}$的凸包中，则对于某个 $k\leqslant n$，当用 u_{n+1} 替换 u_k 时，上述论断仍然成立.

证明　如果可能有 $0=\sum_{i=0}^{n}\theta_i u_i$，其中 $\theta_i\geqslant 0$，$\sum_{i=0}^{n}\theta_i=1$ 及 $\min_i \theta_i=0$. 那么就给出上述表达式并且选取 k 使 $\theta_k=0$，显然，在等式 $0=\sum_{i=0}^{n}\theta_i u_i$ 中我们可以用 u_{n+1} 替换 u_n.

假设不可能有 $0=\sum_{i=0}^{n}\theta_i u_i$，其中 $\theta_i \geqslant 0$，$\sum_{i=0}^{n}\theta_i=1$，$\min_i \theta_i=0$，根据卡拉泰奥多里定理，向量 $u_0,u_1,\cdots,u_n$ 一定生成一个 n 维空间，当然它一定是$\mathbb{R}^n$. 因此可以找到 $\lambda_0,\lambda_1,\cdots,\lambda_n$ 使 $u_{n+1}=\sum_{i=0}^{n}\lambda_i u_i$. 同样，存在 $\theta_i>0$ 使得 $0=\sum_{i=0}^{n}\theta_i u_i$，选取 k 使得对所有 i，都有 $\lambda_k/\theta_k \geqslant \lambda_i/\theta_i$. 那么对于 $0\leqslant i \leqslant n$，令 $\theta'_i=\lambda_k\theta_i-\lambda_i\theta_k$. 而且，令 $\theta'_{n+1}=\theta_k$. 注意到 $\theta'_k=0$. 于是我们有

$$\sum_{i=0}^{n+1}\theta'_i u_i = \sum_{i=0}^{n}\theta'_i u_i+\theta'_{n+1}u_{n+1}=\sum_{i=0}^{n}(\lambda_k\theta_i-\lambda_i\theta_k)u_i+\theta_k u_{n+1}$$
$$=\lambda_k\sum_{i=0}^{n}\theta_i u_i-\theta_k\sum_{i=0}^{n}\lambda_i u_i+\theta_k u_{n+1}$$
$$=\lambda_k(-\theta_k u_k)-\theta_k(u_{n+1}-\lambda_k u_k)+\theta_k u_{n+1}=0$$

此外，因为 $\theta'_{n+1}=\theta_k>0$，可知 $\theta'_i\geqslant 0$，那么对 $i\leqslant n$，

$$\theta'_i=\theta_i\theta_k(\lambda_k/\theta_k-\lambda_i\theta_i)\geqslant 0$$

如果用 $\sum_{i=0}^{n}\theta'_i$ 去除等式 $\sum_{i=0}^{n}\theta'_i u_i=0$，我们可看到0表示成了 $u_0,u_1,\cdots,u_{n+1}$ 的一个凸组合，其中 u_k 没有出现. ■

习题 6.9

1. 求出本节例1中的 c_1，c_2，δ，ξ.
2. 求出一次多项式在区间[0，1]上最佳逼近函数$\sqrt{x}$.
3. 证明：$C[0,1]$中由下列集合生成的子空间是哈尔子空间：
   a. $\{1, x^2, x^3\}$　　b. $\{1, e^x, e^{2x}\}$　　c. $\{(x+2)^{-1}, (x+3)^{-1}, (x+4)^{-1}\}$
4. 证明：$C[-1,1]$中由下列集合生成的子空间不是哈尔子空间：
   a. $\{1, x^2, x^3\}$　　b. $\{|x|, |x-1|\}$　　c. $\{e^x, x+1\}$
5. 在空间 $C[0,1]$中，考虑零次多项式的子空间 Π_0(即常值函数). 利用
   $$M(f)=\max_{0\leqslant x\leqslant 1}f(x) \quad 和 \quad m(f)=\min_{0\leqslant x\leqslant 1}f(x)$$
   描述用 Π_0 中的元在 $C[0,1]$中最佳逼近 f.
6. 设 $u_0,u_1,\cdots,u_n$ 是$\mathbb{R}^n$中 $n+1$ 个点使得不可能有表达式 $0=\sum_{i=0}^{n}\theta_i u_i$，其中 $\theta\geqslant 0,\theta\neq 0$，及$\min_i\theta_i=0$. 证明每个选自$\{u_0,u_1,\cdots,u_n\}$中的 n 个向量的集合是$\mathbb{R}^n$的一组基，或者给出一个反例.
7. 设 A 是 $n\times(n+2)$矩阵. 证明：如果方程组
   $$Ax=0 \quad x\geqslant 0 \quad x\neq 0 \quad x_{n+2}=0 \quad x\in\mathbb{R}^{n+2}$$
   是相容的，那么对某些 $k<n+2$，下面的方程组也相容：
   $$Ax=0 \quad x\geqslant 0 \quad x\neq 0 \quad x_k=0 \quad x\in\mathbb{R}^{n+2}$$
8. 证明：区间[−1，1]上函数 $\cosh x$ 的最佳逼近二次多项式是 $a+bx^2$，其中 $b=\cosh 1-1$ 而 a 可通过求解下面两个同时含 a 和 t 的方程得到：
   $$2a=1+\cosh t-t^2 b$$
   $$\sinh t=2tb$$
9. 证明：一个集合的凸包是凸的而且它还是包含原集合的最小凸集.

419

10. 证明：赋范线性空间中的每个闭球$\{f: \|f-g\| \leqslant r\}$是凸的.
11. 试举出一个凸集的例子，使它的补集有界.
12. 给定平面内一条直线$ax+by+c=0$，其中$a^2+b^2>0$，完整地描述一下与原点距离极小的直线上点的集合，用l_∞范数定义距离.
13. 设f是正方形$0\leqslant x\leqslant 1$，$0\leqslant y\leqslant 1$内(x, y)的连续函数．描述用一个只含x的连续函数最佳逼近f.
14. 下列三个函数
$$g_0(x,y)=1 \quad g_1(x,y)=x \quad g_2(x,y)=y$$
能生成$C(\mathbb{R}^2)$中的一个哈尔子空间吗？
15. 证明：如果在$[a, b]$上$f^{(n)}(x)>0$，那么函数集合$\{1, x, x^2, \cdots, x^{n-1}, f\}$可生成$[a, b]$上的一个哈尔子空间.
16. 设$f=a_0T_0+a_1T_1+\cdots+a_{n+1}T_{n+1}$．其中$T_k$是切比雪夫多项式．证明：空间$\Pi_n$中在区间$[-1, 1]$上采用上确界范数的$f$的最佳逼近是$a_0T_0+a_1T_1+\cdots+a_nT_n$.

6.10 高维插值

寻找多元函数的光滑插值问题是困难的，而这个问题无论是过去还是现在都受到了广泛的关注．多元问题常常会出现一些在一元问题中所没有的异常特征，当仅有两个变量时这些特征就已经很明显了．因此，只讨论二元问题(两个独立变量)几乎不失一般性，至少在最初阶段是这样.

6.10.1 插值问题

我们要讨论的中心问题如下：在xy平面内给定的插值点(或者结点)集合，记为
$$(x_1,y_1) \quad (x_2,y_2) \quad \cdots \quad (x_n,y_n) \tag{1}$$
我们假设这n个点不相同．每个点(x_i, y_i)存在一个对应的实数c_i，而我们的目的是寻找一个光滑并且容易计算的函数F使得

420

$$F(x_i,y_i)=c_i \qquad (1\leqslant i\leqslant n)$$
函数F定义在整个$\mathbb{R}^2$上是不言自明的，或者它至少定义在包含这些结点的某个大的区域上．在以上的描述中，名词光滑和容易计算只具有非正式的或者直观的含义.

6.10.2 笛卡儿积和网格

刚才所描述的插值问题有时可以用一元插值的张量积来求解．我们首先来处理这种情况．此时的结点集记为$\mathcal{N}$，它是**笛卡儿积**
$$\mathcal{N}=\{x_1,x_2,\cdots,x_p\}\times\{y_1,y_2,\cdots,y_q\}$$
因此，$\mathcal{N}$是所有数对(x_i, y_j)的集合，其中x_i选自第一个集合，y_j选自第二个集合．换言之
$$\mathcal{N}=\{(x_i,y_j): 1\leqslant i\leqslant p, 1\leqslant j\leqslant q\} \tag{2}$$
可以看出对这种情况，(2)式中的记号要比(1)式中的记号更方便些．$p=4$和$q=3$时(2)式的一个实例如图6-16所示．结点的这种阵列通常称为**笛卡儿网格**．为方便起见，我们给x点从左往右编号，给y点从下往上编号，当然这并不是必要的.

假设我们有一个结点$x_1, x_2, \cdots, x_p$的线性插值格式．它将是一个**一元过程**．我们要把它看作下列形式的一个线性算子P

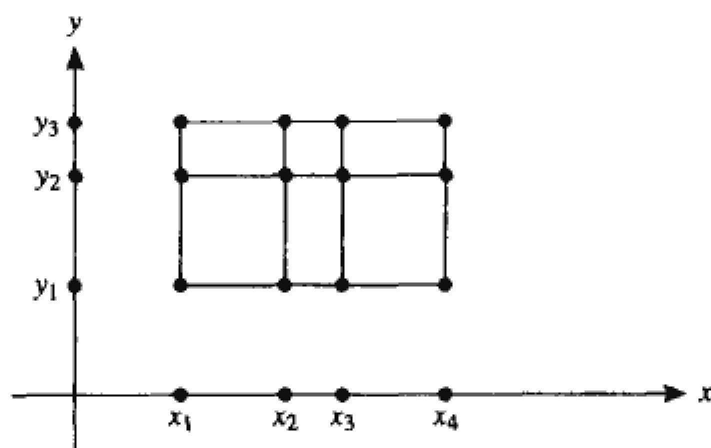


图 6-16 结点的笛卡儿网格

$$(Pf)(x)=\sum_{i=1}^{p} f(x_i)u_i(x) \tag{3}$$

其中函数 u_i 具有基性质

421

$$u_i(x_j)=\delta_{ij} \qquad (1\leqslant i,j\leqslant p) \tag{4}$$

例如，在常规的多项式插值中，函数 u_i 是由 6.1 节中一个熟悉的公式给出的：

$$u_i(x)=\prod_{\substack{j=1\\ j\neq i}}^{p}\frac{x-x_j}{x_i-x_j} \qquad (1\leqslant i\leqslant p) \tag{5}$$

注意到算子 P 可被平凡地推广使其作用在两个或者更多变量的函数上．因此，如果 f 是(x,y)的一个函数，我们可以写出

$$(\overline{P}f)(x,y)=\sum_{i=1}^{p} f(x_i,y)u_i(x) \tag{6}$$

并立即可以看出 $\overline{P}f$ 是在下列垂直线上插值 f 的一个二元函数．

$$L_i\equiv\{(x_i,y):-\infty<y<\infty\} \qquad (1\leqslant i\leqslant p) \tag{7}$$

假设另一个算子可用于结点 y_1，y_2，…，y_q 的插值．我们给出

$$(Qf)(y)=\sum_{i=1}^{q} f(y_i)v_i(y) \tag{8}$$

其中 v_i 是任何适当的函数且具有基性质

$$v_i(y_j)=\delta_{ij} \qquad (1\leqslant i,j\leqslant q) \tag{9}$$

利用下式还可推广 Q 使其作用在二元函数上：

$$(\overline{Q}f)(x,y)=\sum_{i=1}^{q} f(x,y_i)v_i(y) \tag{10}$$

函数 $\overline{Q}f$ 在下列水平线上插值 f：

$$L^i\equiv\{(x,y_i):-\infty<x<\infty\} \qquad (1\leqslant i\leqslant q) \tag{11}$$

6.10.3 布尔和

可利用 $\overline{P}$ 和 $\overline{Q}$ 构造两个有用的二元插值算子：它们是积$\overline{PQ}$和下面定义的布尔和 $\overline{P}\oplus\overline{Q}$，

$$\overline{P}\oplus\overline{Q}=\overline{P}+\overline{Q}-\overline{PQ} \tag{12}$$

根据 $\overline{P}$ 和 $\overline{Q}$ 的定义，很容易推导出这些算子的详细公式. 例如

$$(\overline{PQ}f)(x,y)=\overline{P}(\overline{Q}f)(x,y)=\sum_{i=1}^{p}(\overline{Q}f)(x_i,y)u_i(x)$$

$$=\sum_{i=1}^{p}\sum_{j=1}^{q}f(x_i,y_j)v_j(y)u_i(x) \tag{13}$$

422

因为 $v_j(y_k)u_i(x_l)=\delta_{jk}\delta_{il}$，我们不难看出 $\overline{PQ}f$ 是在所有结点 (x_i, y_i) 上插值 f 的函数. 对算子 $\overline{PQ}$ 也使用张量积的符号 $P\otimes Q$.

以同样的方式可导出 $\overline{P}\oplus\overline{Q}$ 的公式是

$$[(\overline{P}\oplus\overline{Q})f](x,y)=(\overline{P}f)(x,y)+(\overline{Q}f)(x,y)-(\overline{PQ}f)(x,y) \tag{14}$$

$$=\sum_{i=1}^{p}f(x_i,y)u_i(x)+\sum_{j=1}^{q}f(x,y_j)v_j(y)$$

$$-\sum_{i=1}^{p}\sum_{j=1}^{q}f(x_i,y_i)u_i(x)v_j(y)$$

函数 $(\overline{P}\oplus\overline{Q})f$ 在所有水平线 $L_i(1\leqslant i\leqslant p)$ 和所有垂直线 $L^j(1\leqslant j\leqslant q)$ 上插值 f，它的证明留作习题 6.10.6.

例 1 给出一个二元多项式的表达式，它的取值如下：

(x, y)	(1, 1)	(2, 1)	(4, 1)	(5, 1)	(1, 3)	(2, 3)	(4, 3)	(5, 3)	(1, 4)	(2, 4)	(4, 4)	(5, 4)
$f(x, y)$	1.7	−4.1	−3.2	4.9	6.1	−4.2	2.3	7.5	−5.9	3.8	−1.7	2.5

解 首先看出这些结点构成一个笛卡儿网格，因而可应用张量积方法. 由(5)式给出了函数 u_i 和 v_i. 在这例题中，它们是

$$u_1(x)=\frac{x-2}{1-2}\cdot\frac{x-4}{1-4}\cdot\frac{x-5}{1-5}=-\frac{1}{12}(x-2)(x-4)(x-5)$$

$$u_2(x)=\frac{1}{6}(x-1)(x-4)(x-5)$$

$$u_3(x)=-\frac{1}{6}(x-1)(x-2)(x-5)$$

$$u_4(x)=\frac{1}{12}(x-1)(x-2)(x-4)$$

$$v_1(y)=\frac{y-3}{1-3}\cdot\frac{y-4}{1-4}=\frac{1}{6}(y-3)(y-4)$$

$$v_2(y)=-\frac{1}{2}(y-1)(y-4)$$

$$v_3(y)=\frac{1}{3}(y-1)(y-3)$$

因而，多项式插值是

$$\begin{aligned}F(x,y)=&u_1(x)[1.7v_1(y)+6.1v_2(y)-5.9v_3(y)]\\&+u_2(x)[-4.1v_1(y)-4.2v_2(y)+3.8v_3(y)]\\&+u_3(x)[-3.2v_1(y)+2.3v_2(y)-1.7v_3(y)]\\&+u_4(x)[4.9v_1(y)+7.5v_2(y)+2.5v_3(y)]\end{aligned} \tag{15}$$

423

■

6.10.4 张量积

如果把例1中的函数 F 表示成项 x^iy^j 的和，那么会出现下列12项：

$$1,x,x^2,x^3,y,xy,x^2y,x^3y,y^2,xy^2,x^2y^2,x^3y^2 \tag{16}$$

这样，我们用二元多项式的12维子空间进行插值. 这个子空间特有的记号是 $\Pi_3\otimes\Pi_2$. 它是两个线性空间的**张量积**，可由所有如下形式的函数组成：

$$(x,y)\mapsto\sum_{i=1}^{m}a_i(x)b_i(y)$$

其中 $a_i\in\Pi_3$ 及 $b_i\in\Pi_2$.（和式可以有任意多项数.）不难证明（16)式中的函数构成这个空间的一组基.

要强调的一点是刚才概述的理论可应用于一般的函数 u_i 和 v_i，而并不仅仅只应用于多项式. 它们所需要的只是基性质.（在抽象理论中，可以直接利用算子 P 和 Q 进行讨论；而有关它们的详细结构将不记入分析过程.）

在多项式插值的张量积方法中，一般情况将涉及空间 $\Pi_{p-1}\otimes\Pi_{q-1}$ 中的二元多项式，其中 p 和 q 是(2)式中所标出的点的个数. 该空间的一组基由下列函数给出

$$(x,y)\mapsto x^iy^j\qquad(0\leqslant i\leqslant p-1,0\leqslant j\leqslant q-1) \tag{17}$$

因而空间中一般元的形式是

$$(x,y)\mapsto\sum_{i=0}^{p-1}\sum_{j=0}^{q-1}c_{ij}x^iy^j$$

项 x^iy^j 的**次数**定义为 $i+j$. 从而，空间 $\Pi_{p-1}\otimes\Pi_{q-1}$ 中将含有一个次数为 $p+q-2$ 的基元素；即 $x^{p-1}y^{q-1}$. 但是它不包含所有 $p+q-2$ 次的项，例如，不会出现 x^py^{q-2} 项. 一个(x,y)的多项式的次数定义为多项式中各项次数的最大值. 所有次数至多是 k 次的二元多项式空间记为 $\Pi_k(\mathbb{R}^2)$. $\Pi_k(\mathbb{R}^2)$中典型元素是如下形式的函数

$$(x,y)\mapsto\sum_{i=0}^{k}\sum_{j=0}^{k-i}c_{ij}x^iy^j=\sum_{0\leqslant i+j\leqslant k}c_{ij}x^iy^i \tag{18}$$

定理1(二元多项式的基函数定理) $\Pi_k(\mathbb{R}^2)$的一组基是函数集合

424

$$(x,y)\mapsto x^iy^j\qquad(0\leqslant i+j\leqslant k)$$

证明 显然这集合生成 $\Pi_k(\mathbb{R}^2)$，只需要证明它的线性无关性即可. 因此，假设(18)式中函数是0. 如果 y 取定一个固定值，例如 $y=y_0$，那么等式

$$\sum_{i=0}^{k}\Big(\sum_{j=0}^{k-i}c_{ij}y_0^j\Big)x^i=0$$

显示出函数 $x\mapsto x^i$ 之间明显的线性关系. 因为这组函数线性无关，所以我们得到

$$\sum_{j=0}^{k-i}c_{ij}y_0^j=0\qquad(0\leqslant i\leqslant k)$$

在上面等式中，y_0 可以取任意一点. 根据函数组

$$y\mapsto y^j\qquad(0\leqslant j\leqslant k)$$

的线性无关性，我们推断出对所有 i 和 j，有 $c_{ij}=0$ ■

推论 1(二元多项式推论) $\Pi_k(\mathbb{R}^2)$的维数是$(1/2)(k+1)(k+2)$.

证明 定理 1 中给出的$\Pi_k(\mathbb{R}^2)$的基元素可排列如下：

$$\begin{array}{lllll} x^k & & & & \\ x^{k-1} & x^{k-1}y & & & \\ x^{k-2} & x^{k-2}y & x^{k-2}y^2 & & \\ \vdots & \vdots & \vdots & \ddots & \\ x^0 & x^0y & x^0y^2 & \cdots & x^0y^k \end{array}$$

从而基元素的个数是

$$1+2+3+\cdots+(k+1)=\frac{1}{2}(k+1)(k+2)$$ ■

回顾一元多项式的情况，Π_k 可以用于$\mathbb{R}$中任意 $n+1$ 个结点集合上的插值. 很自然地，我们对二元多项式期望$\Pi_k(\mathbb{R}^2)$也可以用于任意 $n\equiv(1/2)(k+1)(k+2)$个结点集合上的插值. 然而，这个期望不会实现，一个简单的例子将说明这一点. 假设 $k=1$，这样 $n=3$. $\Pi_1(\mathbb{R}^2)$中的一般元具有形式

$$c_0+c_1x+c_2y$$

假如我们要求解三个结点(x_i, y_i)上的插值问题，那么我们就要求解一个线性方程组，它的系数行列式是

$$\begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix}$$

425

因为这个行列式表示顶点是给定结点的三角形面积的两倍，所以当这些结点共线时，行列式是零. 因此这种情况的插值问题(一般而言)是不能解的.

6.10.5 几何图形

刚才所考虑的问题表明结点集$\mathcal{N}$的几何图形将决定是否可能在$\mathcal{N}$上用$\Pi_k(\mathbb{R}^2)$插值. 当然，结点个数应该是 $n=(1/2)(k+1)(k+2)$. 这里将给出与该问题相关的几个定理来表明已知的结论. Gasca and Maeztu[1982]给出下述定理.

定理 2(Gasca 和 Maeztu 定理) 若$(1/2)(k+1)(k+2)$个结点集合位于直线 $L_0, L_1, \cdots, L_k$ 上并且(对每个 i)L_i 恰好包含 $i+1$ 个结点，则在子空间$\Pi_k(\mathbb{R}^2)$中可以对这些结点上的任意数据进行插值.

证明 设$\mathcal{N}$表示结点集. 假设定理的前提条件成立，我们有$\#(\mathcal{N}\cap L_i)=i+1$，其中符号#表示集合中元素的个数. 集合$\mathcal{N}\cap L_i$一定两两不相交；否则，将会出现下列矛盾

$$\#\mathcal{N}<\sum_{i=0}^{k}\#(\mathcal{N}\cap L_i)=\sum_{i=0}^{k}(i+1)=\frac{1}{2}(k+1)(k+2)$$

因为结点个数等于空间$\Pi_k(\mathbb{R}^2)$的维数，所以只要证明齐次插值问题只有 0 解就够了. 因此，设 $p\in\Pi_k(\mathbb{R}^2)$，并且假设对$\mathcal{N}$中每一点(x, y)都有 $p(x, y)=0$. 对每个 i，设 ℓ_i 是描述 L_i 的线性函数：

$$L_i = \{(x,y): \ell_i(x,y) = 0\} \qquad (0 \leqslant i \leqslant k)$$

注意到 $p^2+\ell_k^2$ 至少有 $k+1$ 个零点；即 $\mathcal{N} \cap L_k$ 中的点. 由贝祖定理(后面给出)知，ℓ_k 是 p 的因式. 因为 $(p/\ell_k)^2+\ell_{k-1}^2$ 至少有 k 个零点，并且 ℓ_{k-1} 一定是 p/ℓ_k 的因式，故而可以重复以上讨论过程. 在第 k 步之后，我们得知 $\ell_1\ell_2\cdots\ell_k$ 整除 p. 因为 p 的次数至多是 k 次，所以 p 是 $\ell_1\ell_2\cdots\ell_k$ 的一个数量倍数. 因为 p 在 $\mathcal{N} \cap L_0$ 上是零而 $\ell_1\ell_2\cdots\ell_k$ 不是零，所以 p 一定是 0. ■

贝祖定理指出：若 $p\in\Pi_k(\mathbb{R}^2)$，$q\in\Pi_m(\mathbb{R}^2)$，并且 p^2+q^2 的零点数多于 km，那么 p 和 q 一定有非常数的公共因式. 因为这个定理限制在$\mathbb{R}^2$中，所以也适用于定理 2. 后面将给出一个定理 2 的算法证明，其中并不需要用到贝祖定理.

426　**例 2**　图 6-17 所示结点集适合空间 $\Pi_2(\mathbb{R}^2)$中的插值.

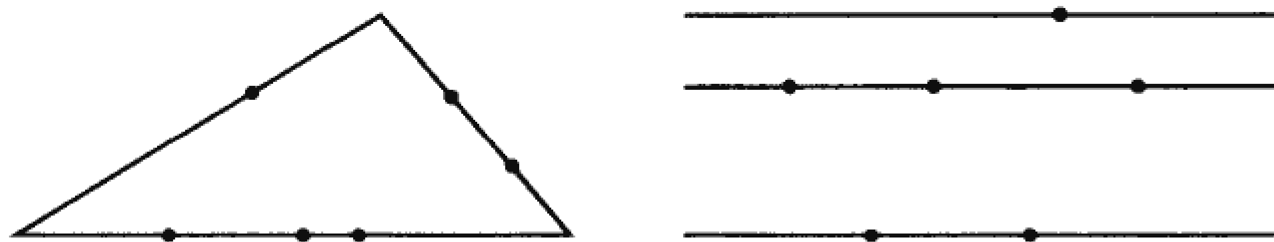


图 6-17　$\Pi_2(\mathbb{R}^2)$中的插值结点集 ■

在高维空间$\mathbb{R}^d$中，可给出与定理 2 密切相关的一个定理. 我们借助习题 6.10.5 得知 $\Pi_k(\mathbb{R}^d)$的维数是$\binom{d+k}{k}$. Chung and Yao[1977]给出下述定理.

定理 3(Chung 和 Yao 定理)　*给定 k 和 d，取 $n=\binom{d+k}{k}$，并且给定$\mathbb{R}^d$中 n 个结点 z_1，z_2，…，z_n. 若存在$\mathbb{R}^d$中的超平面 H_{ij}，其中 $1\leqslant i\leqslant n$ 和 $1\leqslant j\leqslant k$，使得*

$$z_j \in \bigcup_{v=1}^{k} H_{iv} \Leftrightarrow j \neq i \qquad (1 \leqslant i,j \leqslant n) \tag{19}$$

则可用 $\Pi_k(\mathbb{R}^d)$中的多项式插值这个结点集上的任意数据.

证明　每个超平面都是非零线性函数的零点集，因此我们记

$$H_{ij} = \{z \in \mathbb{R}^d: \ell_{ij}(z) = 0\}$$

其中 $\ell_{ij}\in\Pi_1(\mathbb{R}^d)$. 定义函数

$$q_i(z) = \prod_{j=1}^{k} \ell_{ij}(z) \qquad (1 \leqslant i \leqslant n)$$

于是，由条件(19)知 z_i 不属于超平面 H_{i1}，H_{i2}，…，H_{ik} 中的任何一个，所以对 $1\leqslant j\leqslant k$ 有$\ell_{ij}(x_i)\neq 0$. 这就证明了 $q_i(z_i)\neq 0$.

再根据条件(19)，假如 $j\neq i$，那么对某个 v 有 $z_j\in H_{iv}$，因此 $\ell_{iv}(z_j)$和 $q_i(z_j)$都为 0. 这样，如果我们令 $p_i(z)=q_i(z)/q_i(z_i)$，将有基性质 $p_i(z_j)=\delta_{ij}$. 因为 $p_i\in\Pi_k(\mathbb{R}^n)$，从而我们得到用 k 次多项式在这些结点上插值函数 f 的拉格朗日型公式：

$$P(z) = \sum_{i=1}^{n} f(z_i)p_i(z)$$ ■

一个满足定理 3 假设条件的结点形状如图 6-18 所示．其中维数 $d=2$，次数 $k=2$，结点个数 $n=6$．

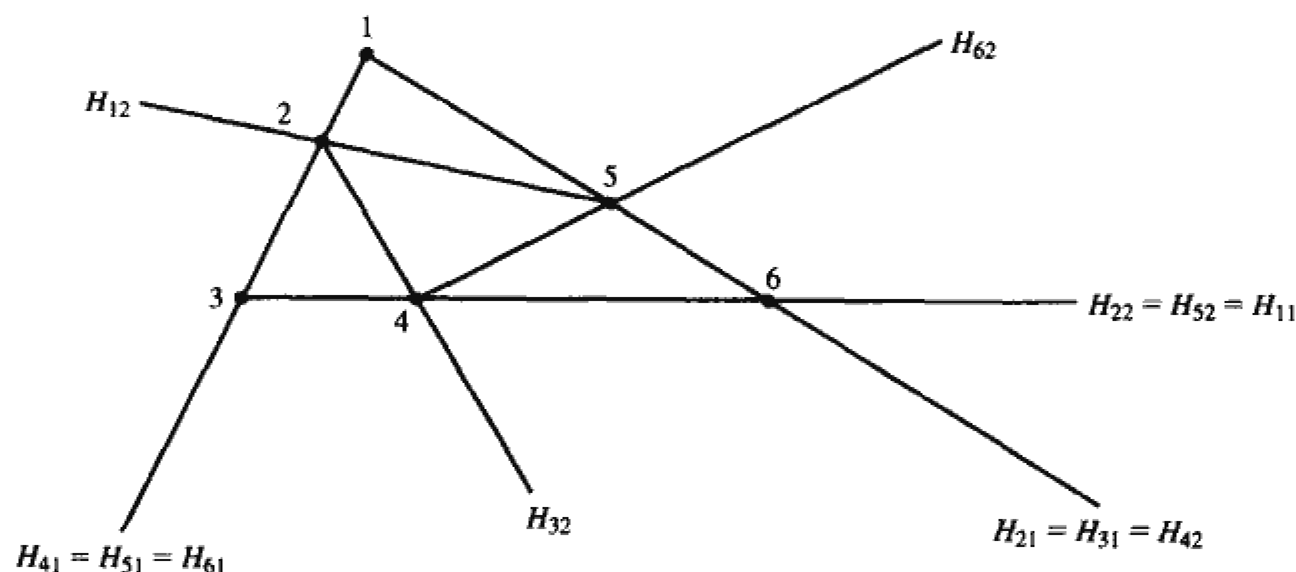


图 6-18 定理 3 的图解

涉及任意结点集上多项式插值的一个很弱的结果如下．

定理 4(插值任意数据的可能性定理) 空间 $\Pi_k(\mathbb{R}^2)$ 是可以对 $\mathbb{R}^2$ 中任一 $k+1$ 个不同结点集上的任意数据插值的．

证明 假设结点是(x_i, y_i)，其中 $0\leqslant i\leqslant k$．我们选择线性函数

$$\ell(x,y)=ax+by+c$$

使得 $k+1$ 个数 $t_i=\ell(x_i, y_i)$都不相同．（见习题 6.10.16.）如果 f 是被插值函数，我们找到 $p\in\Pi_k(\mathbb{R})$，使得 $p(t_i)=f(x_i, y_i)$，那么 $p\circ\ell\in\Pi_k(\mathbb{R}^2)$而且

$$(p\circ\ell)(x_i,y_i)=p(\ell(x_i,y_i))=p(t_i)=f(x_i,y_i)$$ ■

形如 $f\circ\ell$ 的函数，其中 $\ell\in\Pi_1(\mathbb{R}^2)$，被称为**岭函数**．因为 $f\circ\ell$ 在每条直线 $\ell(x, y)=\lambda$ 上是常数，所以它的图形是直纹曲面．定理 4 可直接推广到$\mathbb{R}^d$．

6.10.6 牛顿格式

对于任何一种插值方法，在实际实施过程中，具有像一元多项式插值的牛顿过程那样的算法是有好处的．记得牛顿格式的一个特点是从在结点 x_1，x_2，…，x_n 上插值 f 的多项式 p 出发，通过给 p 添加一项，我们很容易得到在结点 x_1，x_2，…，x_{n+1} 上插值 f 的多项式 p^*．实际上，我们取

$$q(x)=(x-x_1)(x-x_2)\cdots(x-x_n)$$
$$p^*(x)=p(x)+cq(x)$$
$$c=[f(x_{n+1})-p(x_{n+1})]/q(x_{n+1})$$

这个算法的好处是可逐步构造插值多项式，每一步都增加一个新的插值结点并且给 p 增加新的一项．

这个方法的抽象形式如下：设 X 是一个集合及 f 是定义在 X 上的一个实值函数．设 $\mathcal{N}$ 是结点集．如果 p 是 $\mathcal{N}$ 上任一插值 f 的函数，而且 q 是任一在 $\mathcal{N}$ 上取值为零的函数，假如 $q(\xi)\neq 0$，则表达式 $p^*=p+cq$ 给出了在 $\mathcal{N}\cup\{\xi\}$上插值 f 的一个函数 p^*．

427～428

这个方法更一般的形式与结点集有关. 设 q 是 X 到$\mathbb{R}$ 的函数. 并且 Z 是它的 0 点集. 若在$\mathcal{N}\cap Z$ 上 p 插值 f，并且在$\mathcal{N}\setminus Z$ 上 r 插值$(f-p)/q$，则在$\mathcal{N}$上 $p+qr$ 插值 f.

刚才所概述的方法可以用来给出定理 2 的一个算法证明. 首先，选取 $p_k\in\Pi_k(\mathbb{R}^2)$，它在$\mathcal{N}\cap L_k$ 上插值 f. (用定理 4.)我们用向下归纳继续证明. 假设已经在 $\Pi_k(\mathbb{R}^2)$中找到 p_i，它在 $L_k\cup L_{k-1}\cup\cdots\cup L_i$ 中的所有结点上插值 f. 我们试图构造 p_{i-1}具有**牛顿形式**

$$p_{i-1}=p_i+r\ell_k\ell_{k-1}\cdots\ell_i$$

因为给 p_i 的添加项在 $L_k\cup L_{k-1}\cup\cdots\cup L_i$ 中的结点上取值零，显然 p_{i-1}仍然在这集合中的结点上插值 f. 为了使 p_{i-1}在 L_{i-1}中的结点上插值 f，我们记

$$f(x)=p_i(x)+r(x)(\ell_k\ell_{k-1}\cdots\ell_i)(x)\qquad x\in\mathcal{N}\cap L_{i-1}$$

由此我们推断出 r 在$\mathcal{N}\cap L_{i-1}$上插值$(f-p_i)/(\ell_k\ell_{k-1}\cdots\ell_i)$. 根据定理 4 知，存在 $r\in\Pi_{i-1}(\mathbb{R}^2)$满足上述要求. 最后，因为 r 是 $i-1$ 次的，并且$(\ell_k\ell_{k-1}\cdots\ell_i)$是 $k-i+1$ 次的，所以可以看出 $p_{i-1}\in\Pi_k(\mathbb{R}^2)$. 这个算法是 Micchelli[1986a]给出的.

一个有趣的事实是：$C(\mathbb{R}^2)$中根本没有 n 维子空间适合在任意 n 个结点的集合上作插值($n=1$ 的平凡情况除外). 也许是哈尔在 1918 年第一次注意到了这个事实，他是这样论证的. 假设给定 $C(\mathbb{R}^2)$中 n 个函数 u_1，u_2，…，u_n，并且给定$\mathbb{R}^2$中 n 个结点，记为 $p_i=(x_i，y_i)$. 如果我们要用基函数 u_i 在这些结点上作插值，我们就不得不求解一个线性方程组，其系数行列式是

$$D=\begin{vmatrix}u_1(p_1) & u_2(p_1) & \cdots & u_n(p_1)\\ u_1(p_2) & u_2(p_2) & \cdots & u_n(p_2)\\ \vdots & \vdots & \ddots & \vdots\\ u_1(p_n) & u_2(p_n) & \cdots & u_n(p_n)\end{vmatrix}$$

对于给定的结点集，这个行列式可能是非零的. 如果，我们让前两个结点在$\mathbb{R}^2$中作连续的移动，同时在移动的过程中这两个点决不能重合，它们也不能和其他结点重合，那么移动的最终结果是这两个结点相互交换了最初的位置. 根据行列式法则，行列式 D 要变号(因为互换了行列式的第一行与第二行). 由连续性知，在上述连续移动过程中 D 取到值 0. 因此，即使对不重合的结点，D 有时也会等于 0. 在$\mathbb{R}^2$中可以移动两个结点并不经重合就互换位置的事实是$\mathbb{R}^2$,$\mathbb{R}^3$，…所具备的特点，但是$\mathbb{R}^1$可没有这种特点. 这说明了在$\mathbb{R}^2$，$\mathbb{R}^3$，…中处理插值为什么必须有点不同于$\mathbb{R}^1$中的插值. 通常的做法是首先固定结点，然后再询问哪些插值函数的子空间是合适的.

429

6.10.7 Shepard 插值

刚才所述插值类型(其中子空间与结点有关)的一个很一般的方法称为 **Shepard 插值**，它的创始人是 D. Shepard[1968]. 设(不同的)结点如下：

$$p_i=(x_i,y_i)\qquad(1\leqslant i\leqslant n)\tag{20}$$

因为要把相关概念清晰地推广到$\mathbb{R}^3$，$\mathbb{R}^4$，…，所以我们将用 p 和 q 表示$\mathbb{R}^2$中的一般元素. 接下来，我们选取$\mathbb{R}^2\times\mathbb{R}^2$上的一个实值函数 ϕ，使其服从唯一条件

$$\phi(p,q)=0\quad\text{当且仅当}\quad p=q\tag{21}$$

出现在脑海中的例子是 $\phi(p, q)=\|p-q\|$ 及 $\phi(p, q)=\|p-q\|^2$. 然后，我们建立一些完全类似于一元逼近中拉格朗日公式的基函数. 做法如下

$$u_i(p)=\prod_{\substack{j=1\\j\neq i}}^{n}\frac{\phi(p,p_j)}{\phi(p_i,p_j)}\qquad(1\leqslant i\leqslant n)\tag{22}$$

显而易见这些函数具有基性质

$$u_i(p_j)=\delta_{ij}\qquad(1\leqslant i,j\leqslant n)$$

这是(21)式中假设的一个推论. 由此可得

$$F=\sum_{i=1}^{n}f(p_i)u_i\tag{23}$$

是给定结点上插值 f 的函数.

例 3 当 $\phi(p, q)=\|p-q\|^2$ 时，Shepard 插值公式是什么?

解 设 $p_i=(x_i, y_i)$, $p=(x, y)$ 以及 $\phi(p, p_j)=\|p-p_j\|^2=(x-x_j)^2+(y-y_j)^2$. 从而

$$F(x,y)=\sum_{i=1}^{n}f(x_i,y_i)\prod_{\substack{j=1\\j\neq i}}^{n}\frac{(x-x_j)^2+(y-y_j)^2}{(x_i-x_j)^2+(y_i-y_j)^2}$$ ■

例 3 中计算 $F(x, y)$ 的算法如下：

```
input n, x, y
input (x_i, y_i, c_i) (1≤i≤n)
for i=1 to n do
  d_i = (x - x_i)^2 + (y - y_i)^2
  for j=1 to n do
    d_ij = (x_i - x_j)^2 + (y_i - y_j)^2
  end do
end do
```

430

从而

$$F(x,y)=\sum_{i=1}^{n}c_i\prod_{\substack{j=1\\j\neq i}}^{n}d_j/d_{ij}$$

(还没有尝试如何使算法效率更高. 此算法中含有大量的重复工作.)

Shepard 方法还有另一种变形. 首先，附加 ϕ 是非负函数这个条件. 其次，设

$$v_i(p)=\prod_{\substack{j=1\\j\neq i}}^{n}\phi(p,p_j)\qquad v(p)=\sum_{i=1}^{n}v_i(p)\qquad w_i(p)=v_i(p)/v(p)\tag{24}$$

根据 ϕ 的附加条件,如果 $i\neq j$ 并且对除 $p_1,\cdots,p_{i-1},p_{i+1},\cdots,p_n$ 之外的所有点 $v_i(p)>0$,我们有 $v_i(p_j)=0$. 从而有 $v(p)>0$ 而且 w_i 有定义. 由函数的构造可知,$w_i(p_j)=\delta_{ij}$ 以及 $0\leqslant w_i(p)\leqslant 1$. 此外,$\sum_{i=1}^{n}w_i(p)=1$. 下列等式给出了插值过程

$$F=\sum_{i=1}^{n}f(p_i)w_i=\sum_{i=1}^{n}f(p_i)v_i/v\tag{25}$$

这一过程具有前一种形式所不具备的两大优点：即如果数据是非负的，那么插值 F 是非负函数，而且如果 f 是常值函数，则 $F=f$. 这两条性质表明，插值 F 继承了被插值函数的某些特征. 另一方面，如果 ϕ 可微，那么 F 在每个结点上都呈现出一个平坦点. 这是因为 $0\leqslant w_i\leqslant 1$ 及 $w_i(p_j)=\delta_{ij}$，使得结点是每个 w_i 的极点(极大点或者极小点). 从而在每个结点处 w_i 的偏导数为 0，因此，F 也具有同样的结论.

当函数 ϕ 是欧几里得距离的方幂：

$$\phi(x,y) = \| x - y \|^{\mu} \qquad (\mu > 0)$$

其中 x 和 y 是 $\mathbb{R}^s$ 中的点，这时会出现 Shepard 插值的一种重要情况，我们将会看到 $\mu>1$ 时 ϕ 可微，而 $0<\mu\leqslant 1$ 时它不可微. 只需要在可疑点 $x=0$ 处检验简单函数 $g(x)=\|x\|^{\mu}$ 就够了. 通过函数 $G(t)=\|tu\|^{\mu}$ 的微分可得到 g 在点 0 处的方向导数，其中 u 是定义方向的单位向量. 因为 $G(t)=|t|^{\mu}$，所以当 $0<\mu\leqslant 1$ 时，它在点 $t=0$ 处导数不存在，而对 $\mu>1$，$G'(0)=0$.

w_i 的公式可有下面两种形式：

$$w_i(x) = \prod_{\substack{j=1\\ j\neq i}}^{n} \| x - x_j \|^{\mu} \Big/ \sum_{k=1}^{n} \prod_{\substack{j=1\\ j\neq k}}^{n} \| x - x_j \|$$

431

$$w_i(x) = \| x - x_i \|^{-\mu} \Big/ \sum_{j=1}^{n} \| x - x_j \|^{-\mu}$$

由于第二个方程的右端在点 x_i 处呈现不定型∞/∞，所以要谨慎使用.

为了使单个结点上的数据对插值函数在远离结点的点上有较小的影响，Franke 和 Little 设计了*局部多元插值方法*. 给定结点(x_i, y_i)，$1\leqslant i\leqslant n$，我们引入函数

$$g_i(x,y) = \left(1 - r_i^{-1}\sqrt{(x-x_i)^2+(y-y_i)^2}\right)_+^{\mu}$$

下标+表示当括号内是负值时，用 0 代替该负值. 当(x, y)远离结点(x_i, y_i)时就会出现这种情况. 参数 μ 影响函数的光滑性. 参数 r_i 控制 g_i 的支撑集. 从而，假如(x, y)距离(x_i, y_i)超过 r 个单位，则 $g_i(x, y)=0$.

如果选择 r_i 是(x_i, y_i)到其最邻近结点的距离，那么 $g_i(x_j, y_j)=\delta_{ij}$. 这时，可用下列函数插值任意函数 f

$$\sum_{i=1}^{n} f(x_i,y_i)g_i(x,y)$$

6.10.8 三角剖分

从构建**三角剖分**开始，给出在$\mathbb{R}^2$上插值函数的另外一个基本方法. 通俗地说，就是连接结点画出三角形，最终有一族三角形 T_1，T_2，$\cdots T_m$. 我们把这些三角形的配置看作三角剖分. 它们必须满足以下法则：

1. 每个插值结点必须是某个三角形 T_i 的顶点.
2. 配置中三角形的每个顶点一定是结点.
3. 如果一个结点属于一个三角形，那么它一定是那个三角形的顶点.

法则 3 的要求是不允许出现图 6-19 所示的结构.

432

三角剖分上最简单的插值类型是分段线性函数，它在所有三角形的所有顶点上插值函数

f. 在任意三角形 T_i 上，规定线性函数为

$$\ell_i(x,y) = a_i x + b_i y + c_i \qquad (x,y) \in T_i$$

ℓ_i 的系数由 T_i 顶点上给定的函数值唯一确定. 这可以看成是定理 2 的一个应用，因为定理中的 L_1 可选取三角形的一边，而 L_0 可以看作是与 L_1 平行的一条直线并且包含不在 L_1 上的那个顶点.

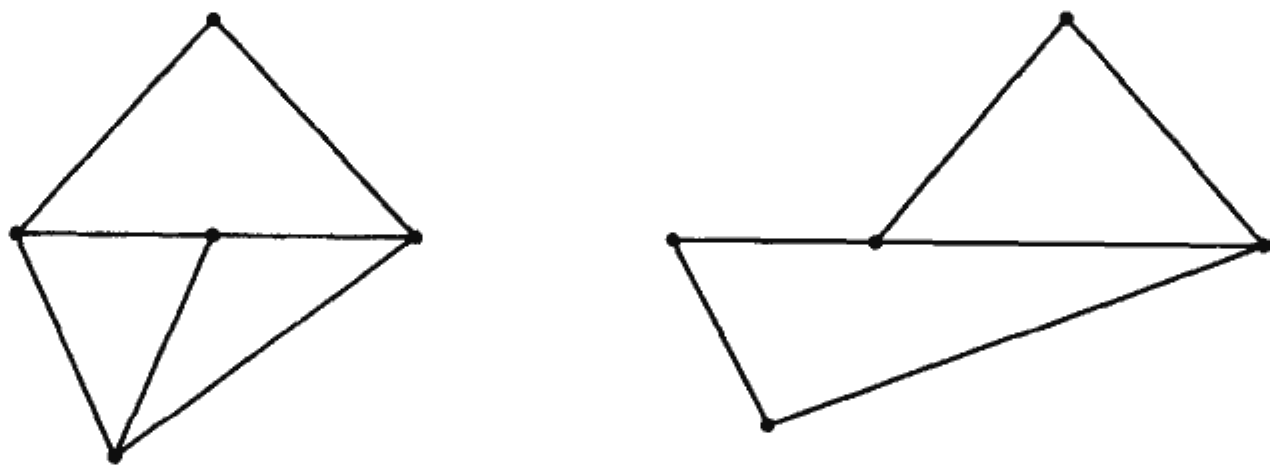


图 6-19 不合规定的三角剖分

考虑一下图 6-20 所示的情况. 连接(x_2, y_2)与(x_3, y_3)的线段是两个三角形的公共边. 这条线段可表示为

$$\{t(x_2,y_2)+(1-t)(x_3,y_3) : 0 \leqslant t \leqslant 1\}$$

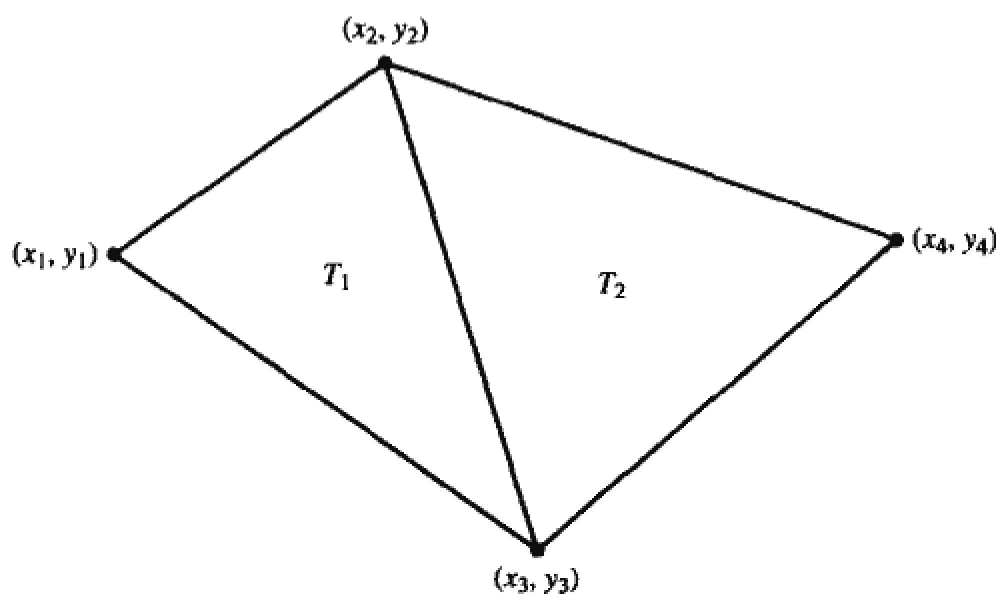


图 6-20 三角剖分

变量 t 可以看作为线段上点的坐标. 当把线性函数 ℓ_1 限制在这条线段上时，它是单变量 t 的线性函数，即

$$a_1(tx_2+(1-t)x_3)+b_1(ty_2+(1-t)y_3)+c_1$$

或者

$$(a_1x_2-a_1x_3+b_1y_2-b_1y_3)t+(a_1x_3+b_1y_3+c_1)$$

这个 t 的线性函数由(x_2, y_2)和(x_3, y_3)上的插值条件完全确定. 同样的结论也适合线性函数

ℓ_2. 因而 ℓ_1 和 ℓ_2 在这条线段上一致，并且定义在 $T_1 \cup T_2$ 上的分段线性函数是连续的. 这就证明了如下结论.

定理 5(定理) 设 $\{T_1, T_2, \cdots, T_m\}$ 是平面上的三角剖分. 在所有三角形 T_i 的所有顶点上取指定值的分段线性函数是连续的.

433

接下来考虑三角剖分上分段二次函数的使用. 在每个三角形 T_i 上，规定二次函数为

$$q_i(x,y) = a_1x^2 + a_2xy + a_3y^2 + a_4x + a_5y + a_6$$

要确定 6 个系数需要 6 个条件. 三角形顶点及各边中点上的值组成了这样的一组条件. 其次，定理 2 的应用还表明这种插值总是唯一可能的. 实际上，在那个定理中，L_2 可以是三角形的一边，L_1 可以是经过不在 L_2 上的两个中点的直线，而 L_0 是含剩余顶点但不含其他结点的直线.（见图 6-21.）同理，因为三角形一条边上的三个给定的函数值确定了该条边上的一元二次函数，所以，我们可以看出整体分段二次函数是连续的.

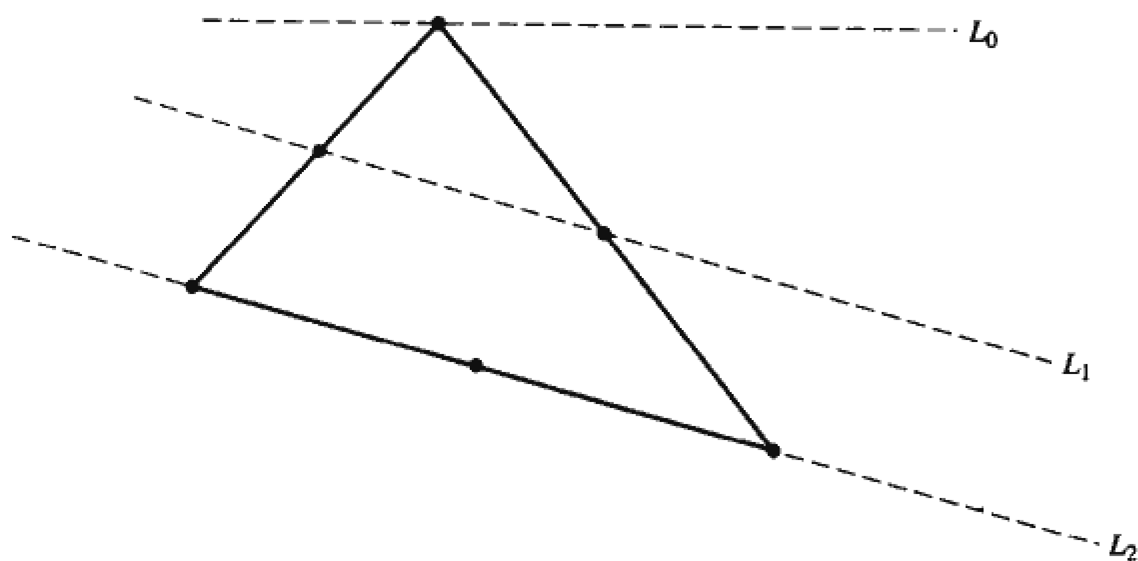


图 6-21 应用定理 2

6.10.9 移动最小二乘法

另一个光滑的并且插值多元函数的通用方法称为**移动最小二乘法**. 首先，我们在一般背景下阐明这种方法，然后给出一些特例.

我们从一个集合 X 开始，它是所涉及函数的定义域. 例如，X 可以是 $\mathbb{R}$，$\mathbb{R}^2$ 或者它们的子集. 其次给定一组结点 $\{x_1, x_2, \cdots, x_n\}$. 某个函数 f 在这些点上已采样. 这样，对 $1 \leqslant i \leqslant n$ 函数值 $f(x_i)$ 已知. 为了进行逼近，我们选取一组函数 $u_1, u_2, \cdots, u_m$. 它们是定义在 X 上的实值函数. 相对于 n 数 m 通常非常小.

在熟知的最小二乘法中，给出了一组非负权 $w_i \geqslant 0$. 我们试图找到系数 $c_1, c_2, \cdots, c_m$ 来极小化表达式

434

$$\sum_{i=1}^{n}\Big[f(x_i) - \sum_{j=1}^{m} c_j u_j(x_i)\Big]^2 w_i$$

这是残差的平方和. 如果我们记

$$\langle f, g\rangle = \sum_{i=1}^{n} f(x_i) g(x_i) w_i \qquad \| f \| = \sqrt{\langle f, f\rangle}$$

那么可应用内积空间中的逼近理论，而且正交性条件

$$f-\sum_{j=1}^{m}c_j u_j \perp u_i \qquad (1\leqslant i\leqslant m)$$

刻画了极小化问题解的特征. 由此式可以导出**正规方程**

$$\sum_{j=1}^{m}c_j\langle u_j,u_i\rangle=\langle f,u_i\rangle \qquad (1\leqslant i\leqslant m)$$

那么移动最小二乘法与刚才概述的过程有何不同呢？此时，允许权 w_i 是 x 的函数. 虽然下列记号可能会更好：

$$\langle f,g\rangle_x=\sum_{i=1}^{n}f(x_i)g(x_i)w_i(x)$$

但是通常的最小二乘法的形式仍然可以保留. 现在正规方程应该记为

$$\sum_{j=1}^{m}c_j(x)\langle u_j,u_i\rangle_x=\langle f,u_i\rangle_x$$

而最终的逼近函数是

$$g(x)=\sum_{j=1}^{m}c_j(x)u_j(x)$$

因为正规方程随 x 一起变化，如果 m 较大，那么产生这个函数所需要的计算将是难以完成的. 因此，m 通常不大于 10.

权函数可用来达到几个预期的效果. 首先，如果 $w_i(x)$ 在 x_i 处是"巨大的"，则函数 g 在 x_i 处几乎插值 f，在极限情况下，$w_i(x_i)=+\infty$ 而且 $g(x_i)=f(x_i)$. 如果当 x 离开 x_i 时 $w_i(x)$ 快速减少到 0，那么远离 x_i 的结点对 $g(x_i)$ 几乎没有影响.

为在空间 $\mathbb{R}^d$ 中实现这两个目标，w_i 的一种选择是

$$w_i(x)=\|x-x_i\|^{-2}$$

其中可使用任何范数，当然欧几里得范数是常用的.

如果移动最小二乘法过程中只用单独一个函数 $u_1(x)\equiv 1$，而且选取刚才所提到的权函数，那么可以导致 Shepard 方法. 为了理解这一点，这时 $c_1(x)=c(x)$，$u_1(x)=u(x)=1$，正规方程记为 435

$$c(x)\langle u,u\rangle_x=\langle f,u\rangle_x$$

逼近函数是

$$g(x)=c(x)u(x)=c(x)=\langle f,u\rangle_x/\langle u,u\rangle_x$$
$$=\sum_{i=1}^{n}f(x_i)w_i(x)\Big/\sum_{j=1}^{n}w_j(x)$$

如果 $w_i(x)=\|x-x_i\|^{-2}$，那么去掉奇点以后，$w_i\Big/\sum_{j=1}^{n}w_j$ 有基性质：它在点 x_i 取值为 1 而在其他结点处取值为 0.

6.10.10 多重二次插值

另一个多元插值过程是 R. L. Hardy 提出的. 它的基函数是所谓的**多重二次函数**

$$z_i(p)=\{\|p-p_i\|^2+c^2\}^{1/2} \qquad (1\leqslant i\leqslant n)$$

其中的范数是欧几里得范数，c 是参数，Hardy 建议 c 等于结点间平均距离的 0.8 倍. 在利用

这些函数的插值过程中，我们需要知道其系数矩阵$(z_i(p_j))$是非奇异的. 这个结论是由Micchelli[1986b]证明的.

有关多元插值方面更多的参考文献是 Chui[1988]、Hartley[1976]、Micchelli[1986a]、Franke[1982] 以及 Lancaster and Salkauskas[1986].

Shepard 插值方法的参考文献是 Shepard[1986]、Gordon and Wixom[1978]、Newman and Rivlin[1983]、Barnhill, Dube, and Little[1983]、Farwing[1986]以及 Cheney and Light[1999].

习题 6.10

1. 给出一个算法用于确定$\mathbb{R}^2$中的点集$\mathcal{N}$是否可以表示成(2)式中的笛卡儿积. 如果可以分解，则算法应该提供因子.
2. 对于下列每个包含关系，或者证明它是对的，或者证明它是错的：
   a. $\Pi_n \otimes \Pi_m \subseteq \Pi_{n+m}(\mathbb{R}^2)$
   b. $\Pi_k(\mathbb{R}^2) \subseteq \Pi_n \otimes \Pi_m$，其中 $k=\max(n, m)$
   c. $\Pi_k(\mathbb{R}^2) \subseteq \Pi_n \otimes \Pi_m$，其中 $k=\min(n, m)$
3. 证明：$\dim \Pi_k(\mathbb{R}^3)=(1/6)(k+1)(k+2)(k+3)$. 其中空间是由次数至多是 k 次的所有三元多项式组成.
4. 本题和下题涉及 d 元多项式. 给每个变元标注下标并把它们放在一个称为 x 的向量中：

436

$$x = \{x_1, x_2, \cdots, x_d\} \in \mathbb{R}^d$$

现在我们需要**多重指标**的概念. 这是一个 d 维的非负整数

$$\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_d\} \in \mathbb{Z}_+^d$$

我们定义

$$x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} x_3^{\alpha_3} \cdots x_d^{\alpha_d}$$

它被称为一个**多项式**并且它是 d 元多项式的基本构件. 多重指标 α 的阶定义为

$$|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_d$$

我们注意到多项式 x^α 的次数恰好是 $|\alpha|$. 证明：$x^\alpha x^\beta = x^{\alpha+\beta}$. 并证明：任意次数$\leqslant k$ 的 d 元多项式可以表示成下列形式：

$$\sum_{|\alpha| \leqslant k} c_\alpha x^\alpha$$

5. 设 $\Pi_k(\mathbb{R}^d)$是所有次数至多是 k 次的 d 元多项式集合. 证明：这个空间的维数是$\binom{d+k}{k}$.
6. 证明课本中(例 1 的前面)有关算子 $\overline{P} \oplus \overline{Q}$ 的插值性质.
7. 如果 6 个结点的集合位于一个椭圆上，或者抛物线上，或者双曲线上，或者一对直线上. 证明：在这个结点集上用 $\Pi_k(\mathbb{R}^2)$插值一般是不可能的.
8. 什么类型的二元多项式适合在图 6-22 所示结点集上的插值？

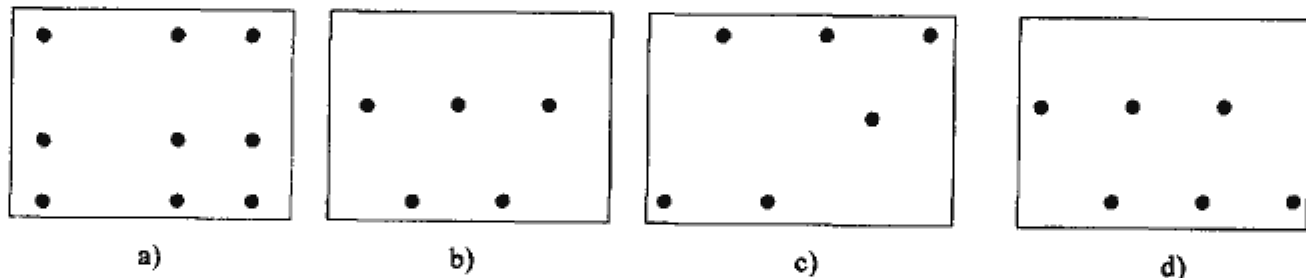


图 6-22

9. 证明：下列函数是 $\Pi_n \otimes \Pi_m$ 的一组基：

$$(x,y) \mapsto x^i y^j \qquad (0 \leqslant i \leqslant n, 0 \leqslant j \leqslant m)$$

10. 利用函数 $\phi(p, q) = \| p-q \|_\infty^2$，给出第二种 Shepard 插值形式的公式．记住，这里的范数是 $\|(x, y)\| = \max\{|x|, |y|\}$.
11. 下列形式的二元多项式

$$F(x,y) = \sum_{0 \leqslant i+j \leqslant k} c_{ij} x^i y^j$$

称为至多 k 次的多项式．如果有一个系数 $c_{ij} \neq 0$ 其中 $i+j=k$，我们称 F 是 k 次的．证明：一个 k 次二元多项式与一个 m 次二元多项式的乘积是一个 $m+k$ 次的多项式.
12. 考虑 n 个结点两个变元的 Shepard 插值，而且函数 ϕ 定义为 $\phi(p, q) = \| p-q \|^2$．试问什么样的多项式空间 $\Pi_k(\mathbb{R}^2)$ 包含所有的基函数？（给出最小的 k.）
13. 设 $f \in \Pi_k(\mathbb{R})$ 及 $\ell \in \Pi_1(\mathbb{R}^2)$．证明：$f \circ \ell \in \Pi_k(\mathbb{R}^2)$．试问 $\Pi_k(\mathbb{R}^2)$ 中每个元都可以用这种方式得到吗？ 437
14. 假设一个矩形的边分别与坐标轴平行，以矩形的顶点为结点．证明：下列多项式可以对这 4 个结点上的任意数据作插值.

$$p(x,y) = a + bx + cy + dxy$$

15. 证明：一个线性多项式

$$p(x,y) = ax + by + c$$

在一个三角形的顶点上总可以作插值．给出两种证明，其中一种证明要基于定理 3.
16. 证明：如果给定 $\mathbb{R}^d$ 中有 n 个不同的点 x_i，那么存在一个向量 b 使得 n 个数 $t_i = \langle x_i, b \rangle$ 都不相同．（定理 4 需要这个结论.）
17. 若 U 和 V 分别是 X 和 Y 上函数的向量空间，则 $U \otimes V$ 由所有函数 w 组成，其中 w 可表示为有限和形式：

$$w(x,y) = \sum_{i=1}^{n} u_i(x) v_i(y) \qquad u_i \in U, v_i \in V$$

证明：若 U 和 V 都是有限维的，则 $\dim(U \otimes V) = \dim U \times \dim V$.
18. 证明：对于算子 L

$$Lf = \sum_{i=1}^{n} f(x_i) w_i$$

下列性质等价：

a. 对所有 f，$\min f(x_i) \leqslant Lf \leqslant \max f(x_i)$.

b. $w_i \geqslant 0$ 并且 $\sum_{i=1}^{n} w_i = 1$.

19. 试问下面的说法是三角剖分法则的一种准确地概括吗？结点集恰好与三角形的顶点集相同.

计算机习题 6.10

编写并测试一个实施 Shepard 方法的有效代码．用户要具体指定结点 (x_i, y_i)，纵坐标 c_i，以及一串点．在这些点上可以求出 Shepard 插值的值．如果有可能，得到图形的输出信息.

6.11 连分式

在数学应用中出现的许多特殊函数都是用无穷过程来定义的，例如级数、积分和迭代等．连分式就是这种无穷过程之一．连分式的一个范例是由 Lambert 于 1770 年给出的：

438

$$\tan^{-1}x = \cfrac{x}{1+\cfrac{x^2}{3+\cfrac{4x^2}{5+\cfrac{9x^2}{7+\cfrac{16x^2}{9+\cdots}}}}} \qquad (|x|<1) \tag{1}$$

它还可以写成

$$\tan^{-1}x = \frac{x}{1+}\,\frac{x^2}{3+}\,\frac{4x^2}{5+}\,\frac{9x^2}{7+}\,\frac{16x^2}{9+}\cdots \qquad (|x|<1) \tag{2}$$

上面等式的右端表示这样的一个极限：在(2)式中第 n 项后终止的表达式给出了一个完全确定的函数 f_n：

$$f_n(x) = \frac{x}{1+}\,\frac{x^2}{3+}\,\frac{4x^2}{5+}\cdots\frac{(n-1)^2x^2}{2n-1} \qquad (n\geqslant 2) \tag{3}$$

这称为连分式的 *n 次渐近分式*. (2)式意味着

$$\tan^{-1}x = \lim_{n\to\infty} f_n(x) \qquad (|x|<1) \tag{4}$$

如果我们假设这等式正确，那么就有另一种计算正切函数值的方法. 该方法是否实用与(4)式中收敛的*速度*有关. 为了用数值方法来判断这一点，对 $n\geqslant 2$ 我们用序列 $f_n(1/\sqrt{3})$ 计算 $\tan^{-1}(1/\sqrt{3})=\pi/6\approx 0.523\,598\,775\,6$. 结果如下：

n	$f_n(1/\sqrt{3})$
2	0.519 615
3	0.523 892
4	0.523 577
5	0.523 600
6	0.523 599
7	0.523 599

这数表显示在第 6 次渐进分式中就可得到 6 位小数的精度.

6.11.1 递归公式

计算连分式并不像计算级数那么容易. 对于一个无穷级数，例如 $\sum_{k=1}^{\infty}a_k$，我们用公式 $S_{n+1}=S_n+a_{n+1}$ 计算部分和，其中 $S_n=\sum_{k=1}^{n}a_k$. 我们对连分式考虑类似的问题：

$$C = \frac{a_1}{b_1+}\,\frac{a_2}{b_2+}\,\frac{a_3}{b_3+}\cdots \tag{5}$$

我们要找到一个递归公式用于计算相继的渐进分式：

$$C_n = \frac{a_1}{b_1+}\,\frac{a_2}{b_2+}\,\frac{a_3}{b_3+}\cdots\frac{a_{n-1}}{b_{n-1}+}\,\frac{a_n}{b_n} \tag{6}$$

为了暂时使用，我们引进函数

$$f_n(x) = \frac{a_1}{b_1 +} \frac{a_2}{b_2 +} \frac{a_3}{b_3 +} \cdots \frac{a_{n-1}}{b_{n-1} +} \frac{a_n}{b_n + x} \quad (7)$$

439

显然，我们得到

$$C_n = f_n(0)$$

为了从 $f_{n-1}(x)$ 得到 $f_n(x)$，首先我们看出

$$f_n(x) = f_{n-1}\left(\frac{a_n}{b_n + x}\right)$$

通过直接计算，有

$$f_1(x) = \frac{a_1}{b_1 + x}$$

$$f_2(x) = \frac{a_1 b_2 + a_1 x}{b_1 b_2 + a_2 + b_1 x}$$

$$f_3(x) = \frac{a_1 b_2 b_3 + a_1 a_3 + (a_1 b_2) x}{b_1 b_2 b_3 + a_2 b_3 + b_1 a_3 + (b_1 b_2 + a_2) x}$$

上面所出现的模式使人联想到

$$f_n(x) = \frac{A_n + A_{n-1} x}{B_n + B_{n-1} x} \qquad (n \geqslant 1) \quad (8)$$

其中

$$\begin{cases} A_0 = 0, A_1 = a_1 \\ A_n = b_n A_{n-1} + a_n A_{n-2} \qquad (n \geqslant 2) \end{cases} \quad (9)$$

以及

$$\begin{cases} B_0 = 1, B_1 = b_1 \\ B_n = b_n B_{n-1} + a_n B_{n-2} \qquad (n \geqslant 2) \end{cases} \quad (10)$$

定理 1（连分式定理） *若给定序列 $\{a_n\}_{n=1}^{\infty}$ 和序列 $\{b_n\}_{n=1}^{\infty}$，而且序列 $\{A_n\}_{n=0}^{\infty}$ 和序列 $\{B_n\}_{n=0}^{\infty}$ 分别由(9)式和(10)式定义，则*

$$C_n = \frac{A_n}{B_n} \qquad (n \geqslant 1) \quad (11)$$

证明 对指标 1，2，3，已经证明了(8)式正确．现在假设对指标 1，2，…，$n-1$，(8)式也成立．于是

440

$$\begin{aligned} f_n(x) &= f_{n-1}\left(\frac{a_n}{b_n + x}\right) \\ &= \frac{A_{n-1} + A_{n-2} a_n/(b_n + x)}{B_{n-1} + B_{n-2} a_n/(b_n + x)} \\ &= \frac{A_{n-1}(b_n + x) + A_{n-2} a_n}{B_{n-1}(b_n + x) + B_{n-2} a_n} \\ &= \frac{A_{n-1} b_n + A_{n-2} a_n + A_{n-1} x}{B_{n-1} b_n + B_{n-2} a_n + B_{n-1} x} \\ &= \frac{A_n + A_{n-1} x}{B_n + B_{n-1} x} \end{aligned}$$

根据数学归纳法知(8)式正确. 那么，当然有

$$C_n = f_n(0) = \frac{A_n}{B_n}$$ ■

这里推导出来的递归公式构成了一个有效算法的基础. 例如，本节一开始给出 $\tan^{-1}(1/\sqrt{3})$ 的数值结果，就是用这个递归公式计算的，从 $a_1=x$、$b_1=1$ 开始，并且 $a_n=(n-1)^2x^2$ 及 $b_n=2n-1$.

6.11.2 级数到连分式的转换

应用数学中许多重要的特殊函数都有连分式展开. 有关信息可参见 Abramowitz and Stegun[1964]，相关的教科书有 Khovanskii[1963]、Perron[1929]以及 Wall[1948]. 我们只在这里简要地说明一个过程，通过它可从级数得到连分式.

定理 2(级数到连分式定理)

$$\sum_{k=1}^{\infty}\frac{1}{x_k} = \frac{1}{x_1-}\,\frac{x_1^2}{x_1+x_2-}\,\frac{x_2^2}{x_2+x_3-}\cdots\frac{x_{n-1}^2}{x_{n-1}+x_n-}\cdots \tag{12}$$

证明 这可以用数学归纳法来证明，我们把它留给读者作为习题 6.11.17. ■

为解释如何使用该定理，我们用 arctan x 的麦克劳林级数构造出一个连分式：

$$\begin{aligned}\arctan x &= x-\frac{x^3}{3}+\frac{x^5}{5}-\frac{x^7}{7}+\cdots\\ &= \frac{1}{x^{-1}}+\frac{1}{-3x^{-3}}+\frac{1}{5x^{-5}}+\frac{1}{-7x^{-7}}+\cdots\\ &= \frac{1}{x^{-1}-}\,\frac{(x^{-1})^2}{x^{-1}+(-3x^{-3})}-\frac{(-3x^{-3})^2}{(-3x^{-3})+(5x^{-5})-}\,\frac{(5x^{-5})^2}{(5x^{-5})+(-7x^{-7})-}\cdots\\ &= \frac{x}{1+}\,\frac{-x^2}{x^2-3+}\,\frac{-9x^2}{-3x^2+5+}\,\frac{-25x^2}{5x^2-7-}\cdots\end{aligned}$$

[441]

我们注意到，如果连分式中某一分量分式的分子和分母都被乘以相同的量，那么下一个分量分式的分子也会受到影响. (为什么?)

对于 $n\geqslant 2$，当

$$a_n=-(2n-3)^2x^2 \quad 和 \quad b_n=(-1)^n[(2n-3)x^2-(2n-1)]$$

对给定的 x 值，(9)式，(10)式和(11)式可用于逼近 arctan x. 不过，利用这个连分式计算 arctan x 的算法要远比基于计算麦克劳林级数部分和的算法复杂很多，而这两个算法产生相同的数列. 可以看出，(1)式中 arctan x 的兰伯特连分式与通过麦克劳林级数推出的连分式是不同的. 为进一步看清这个问题，我们用下列级数的部分和计算 $\arctan(1/\sqrt{3})$：

n	$f_n(1/\sqrt{3})$
1	0.577 350
2	0.513 200
3	0.526 030
4	0.522 976
5	0.523 767

6	0.523 551
7	0.523 612
8	0.523 595
9	0.523 600
10	0.523 598
11	0.523 599
12	0.523 599

把该数表与由连分式(1)产生的数表作比较表明，就这个例子而言，连分式比级数收敛的更快.

习题 6.11

1. 证明：

$$\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\cdots=\frac{1}{2}(\sqrt{5}-1)$$

提示：令 x 等于这个连分式，然后查看 $1/x$.

2.（续）对上题中的连分式，证明：

$$C_n=\frac{A_n}{A_{n+1}}$$

[442]

3. 证明：

$$\sqrt{x}=1+2\left(\frac{v}{1+}\frac{v}{1+}\frac{v}{1+}\cdots\right)$$

其中 $v=(1/4)(x-1)$.

4.（续）对上题中的连分式，证明：

$$C_n=\frac{uA_n}{A_{n+1}}$$

5. 证明：

$$\sqrt{b^2+a}=b+\frac{a}{2b+}\frac{a}{2b+}\frac{a}{2b+}\cdots$$

6. 比较计算下列函数值的两种方法：

$$f(x)=\int_0^x e^{-t^2}\,dt$$

即习题 6.7.6 给出的泰勒级数展开式和连分式

$$f(x)=\frac{\sqrt{\pi}}{2}-\frac{1}{2}e^{-x^2}\left(\frac{1}{x+}\frac{1}{2x+}\frac{2}{x+}\frac{3}{2x+}\frac{4}{x+}\cdots\right)$$

7. 证明：区间 $0<x<1$ 中每个实数都可表示成下列形式的连分式(也许是可中止的)：

$$x=\frac{1}{b_1+}\frac{1}{b_2+}\frac{1}{b_3+}\cdots$$

其中每个 b_i 都是正整数.

8. 证明：如果

$$x=\frac{a_1}{b_1+}\frac{a_2}{b_2+}\frac{a_3}{b_3+}\cdots$$

并且 a_i 和 b_i 都是正的，那么连分式的渐进分式交替地大于 x 和小于 x..

9. 使用(8)式的记号，证明：f_n 是递增的当且仅当 $A_{n-1}B_n>A_nB_{n-1}$.

10. 计算出利用(9)式和(10)式的 A_n/B_n 求 C_n 过程中长运算的个数.

11. 证明：(11)式的连分式中任意两个相继的渐进分式服从方程

$$\frac{A_n}{B_n}-\frac{A_{n-1}}{B_{n-1}}=(-1)^{n-1}\frac{a_1a_2\cdots a_n}{B_nB_{n-1}}$$

12.(续)利用上题的结论证明习题 6.11.8.

13. 证明：如果数 b_i 都是正的并且 $a_i=1$，那么对所有 n 都有 $B_n>\min(1, b_1)$.

14. 证明：如果 $b_i>0$ 并且 $a_i=1$，那么作为 n 的函数 B_nB_{n-1} 应该是递增的.

15. 证明：如果 $b_i\geqslant 1=a_i$，那么(11)式中的连分式收敛.

16. 如果

$$\frac{a_1}{b_1+}\frac{a_2}{b_2+}\cdots=\frac{1}{B_1+}\frac{1}{B_2+}\cdots$$

求出 B_n 的递归公式.

443 17. 证明本节中的定理 2. 提示：用数学归纳法.

18. 证明：

$$e^x=\frac{1}{1-}\frac{x}{x+1-}\frac{x}{x+2-}\frac{2x}{x+3-}\frac{3x}{x+4-}\cdots$$

19.(续)证明：

$$\frac{1}{1+}\frac{2}{2+}\frac{3}{3+}\cdots=\frac{1}{e-1}$$

提示：利用上题.

20. 证明：函数

$$g_n(x)=\frac{1}{x+}\frac{2}{x+}\frac{3}{x+}\cdots\frac{n}{x}$$

可由下列算法递归生成：

$$g_n(x)=\frac{p_n(x)}{q_n(x)}$$

其中

$$\begin{cases} p_0(x)=0, \quad p_1(x)=1 \\ p_{n+1}(x)=xp_n(x)+(n+1)p_{n-1}(x) \end{cases}$$

以及

$$\begin{cases} q_0(x)=1, \quad q_1(x)=x \\ q_{n+1}(x)=xq_n(x)+(n+1)q_{n-1}(x) \end{cases}$$

21. 假设下列连分式收敛，求出它的值：

$$\frac{1}{6+}\frac{1}{6+}\frac{1}{6+}\frac{1}{6+}\cdots$$

22. 求出 x 的值：

$$x=1+\frac{1}{1+}\frac{1}{2+}\frac{1}{1+}\frac{1}{2+}\cdots$$

23. 求出 x 的值：

$$x=2+\frac{1}{4+}\frac{1}{4+}\frac{1}{4+}\cdots$$

24. 如果

$$f_n(x) = \frac{2}{1+}\frac{4}{2+}\frac{6}{3+}\cdots\frac{2n}{n+x}$$

你如何从 f_n 得到 f_{n+1}？

25. 求$\sqrt{6+\sqrt{6+\sqrt{6+\cdots}}}$的值.

26. 假设连分式

$$\frac{1}{2x+}\frac{1}{2x+}\frac{1}{2x+}\cdots \qquad (x>0)$$

收敛. 确定一个它关于 x 的闭型表达式.

计算机习题 6.11

1. 利用习题 6.11.3 中的等式，编写一个计算$\sqrt{x}$的程序. 并通过打印出前 50 项渐进分式值的数表来计算 $\sqrt{10}$，$\sqrt{100}$，$\sqrt{1\,000}$，$\sqrt{10\,000}$.

2. 编写一个计算 $\arctan(1/\sqrt{3})$的程序，不要使用下标变量，并与课文中的结果作比较. [444]

3. 编写一个计算机程序用于计算(1)式中给出的 $\arctan x$ 的连分式. 利用如下基本方法：给定 n，从(3)式右端开始构造出适当的分式来计算其中的 $f_n(x)$. 通过计算 $\pi^{-1}\arctan(\sqrt{3})$来测试你的程序，其中 $n=5$，10，15，20.

6.12 三角插值

首先，我们回顾一下关于通常的代数多项式插值的一些显著的事实. 如果 $n+1$ 个函数值由下表给出：

x	x_0	x_1	x_2	$\cdots$	x_n
y	y_0	y_1	y_2	$\cdots$	y_n

(1)

那么存在唯一一个次数$\leqslant n$ 的多项式 p 插值这些数据. 换言之，

$$p(x_j) = y_j \qquad (0\leqslant j\leqslant n) \tag{2}$$

假设点 x_0，x_1，…，x_n 是相互不同的，但是对数据 y_j 没有作限制. 多项式 p 从由所有次数$\leqslant n$ 的多项式组成的线性空间 Π_n 中选取. Π_n 的一组基由函数 $b_k(x)=x^k$ 给出，$0\leqslant k\leqslant n$.

6.12.1 傅里叶级数

当然，代数多项式空间 Π_n 不适合表示周期现象. 而三角函数却是非常适合的. 在选择基本三角函数之前我们必须知道问题中的函数周期. 为方便起见，我们假设被插值函数是周期为 2π 的周期函数. 于是，选取函数 1，$\cos x$，$\cos 2x$，…及 $\sin x$，$\sin 2x$，…是比较适当的. 傅里叶分析的一个基本定理就指出：若 f 是 2π 周期并且有连续的一阶导数，则傅里叶级数

$$\frac{a_0}{2} + \sum_{k=1}^{\infty}(a_k\cos kx + b_k\sin kx) \tag{3}$$

一致收敛于 f. 该级数中的傅里叶系数可由下列公式计算

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\cos kt\,dt \tag{4}$$

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin kt\,dt \tag{5}$$

[445] 所引用的这个定理使我们确信用上面列出的正弦和余弦函数来逼近 2π 周期函数是合理的.

6.12.2 复傅里叶级数

傅里叶级数理论中的许多内容可用复数指数表示成非常优美的形式. 回顾**欧拉公式**:

$$e^{i\theta} = \cos\theta + i\sin\theta \tag{6}$$

其中 $i^2 = -1$. 傅里叶级数为

$$f(x) \sim \sum_{k=-\infty}^{\infty} \hat{f}(k) e^{ikx} \tag{7}$$

其中

$$\hat{f}(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t) e^{-ikt}\,dt \tag{8}$$

如果 f 是实函数, 那么它在(3)式中的傅里叶级数是(7)式中复傅里叶级数的实部. 事实上, 根据(4)式, (5)式, (6)式和(8)式, 我们有

$$\hat{f}(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)[\cos kt - i\sin kt]\,dt = \frac{1}{2}(a_k - ib_k) \qquad (k \geqslant 0) \tag{9}$$

现在, 我们引出下述定理.

定理 1(傅里叶级数定理) 给定实序列 $[a_k]_{k=0}^{\infty}$ 和 $[b_k]_{k=0}^{\infty}$, 定义

$$b_0 = 0 \qquad a_{-k} = a_k \qquad b_{-k} = -b_k \qquad c_k = \frac{1}{2}(a_k - ib_k)$$

则

$$\frac{a_0}{2} + \sum_{k=1}^{n}(a_k\cos kx + b_k\sin kx) = \sum_{k=-n}^{n} c_k e^{ikx} \tag{10}$$

证明 右端的级数可以写为

$$\frac{1}{2}\sum_{k=-n}^{n}(a_k - ib_k)(\cos kx + i\sin kx) \tag{11}$$

如下列计算所示, 该级数的虚部是 0.

$$\begin{aligned}
&\frac{1}{2}\sum_{k=-n}^{n}[a_k\sin kx - b_k\cos kx] \\
&= \frac{1}{2}\sum_{k=1}^{n}[a_{-k}\sin(-kx) - b_{-k}\cos(-kx)] - \frac{b_0}{2} + \frac{1}{2}\sum_{k=1}^{n}[a_k\sin kx - b_k\cos kx] \\
&= \frac{1}{2}\sum_{k=1}^{n}[-a_k\sin kx + b_k\cos kx] + \frac{1}{2}\sum_{k=1}^{n}[a_k\sin kx - b_k\cos kx] = 0
\end{aligned}$$

[446]

(11)式中级数的实部是

$$\begin{aligned}
&\frac{1}{2}\sum_{k=-n}^{n}[a_k\cos kx + b_k\sin kx] \\
&= \frac{1}{2}\sum_{k=1}^{n}[a_{-k}\cos(-kx) + b_{-k}\sin(-kx)] + \frac{a_0}{2} + \frac{1}{2}\sum_{k=1}^{n}[a_k\cos kx + b_k\sin kx] \\
&= \frac{a_0}{2} + \sum_{k=1}^{n}(a_k\cos kx + b_k\sin kx)
\end{aligned}$$

■

6.12.3 内积，伪内积，伪范数

复希尔伯特空间 $L_2[-\pi, \pi]$中的**内积**定义为

$$\langle f,g\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\,\overline{g(x)}\mathrm{d}x$$

由 $E_k(x)=\mathrm{e}^{ikx}$(其中 $k=0$, ±1, ±2, …)定义的函数 E_k 构成一个复希尔伯特空间中函数的标准正交系. 这意味着当 $n\neq k$ 时$\langle E_k, E_n\rangle=0$ 并且$\langle E_k, E_k\rangle=1$. 这可以从下面的计算中看出. 当 $n\neq k$ 时

$$\langle E_k,E_n\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} E_k(x)\,\overline{E_n(x)}\mathrm{d}x = \frac{1}{2\pi}\int_{-\pi}^{\pi}\mathrm{e}^{ikx}\mathrm{e}^{-inx}\mathrm{d}x$$
$$= \frac{1}{2\pi}\int_{-\pi}^{\pi}\mathrm{e}^{\mathrm{i}(k-n)x}\mathrm{d}x = \frac{1}{2\pi}\frac{\mathrm{e}^{\mathrm{i}(k-n)x}}{\mathrm{i}(k-n)}\Bigg|_{x=-\pi}^{x=\pi} = 0$$

显然地$\langle E_k, E_k\rangle=1$，而且函数 E_k 构成一个标准正交序列.

使用下列内积符号将会很方便

$$\langle f,g\rangle_N = \frac{1}{N}\sum_{j=0}^{N-1} f(2\pi j/N)\,\overline{g(2\pi j/N)} \tag{12}$$

因为从条件$\langle f, f\rangle_N=0$ 不能推得 $f=0$，只能得知在每个结点 $2\pi j/N$ 上函数 $f(x)$的值是 0，所以这个函数不是一个真正的内积. 但是它符合(复)内积的其他性质. 它们是：

复内积性质

1. $\langle f, f\rangle_N\geqslant 0$.
2. $\langle f, g\rangle_N=\overline{\langle g, f\rangle}_N$.
3. $\langle \alpha f+\beta g, h\rangle_N=\alpha\langle f, h\rangle_N+\beta\langle g, h\rangle_N$.

447

与**伪内积**一起，还有下面定义的**伪范数**：

$$\| f \|_N = \sqrt{\langle f,f\rangle_N}$$

我们有 $\| f \|_N=0$ 当且仅当 $f(2\pi j/N)=0$，其中 $0\leqslant j\leqslant N-1$.

对于插值而言，下述定理起着决定性的作用.

定理 2(伪内积定理) 对任意 $N\geqslant 1$，我们有

$$\langle E_k,E_m\rangle_N = \begin{cases}1 & 若\ N\ 整除\ k-m\\ 0 & 其他\end{cases} \tag{13}$$

证明 问题中的表达式可写为

$$\frac{1}{N}\sum_{j=0}^{N-1}E_k\left(\frac{2\pi j}{N}\right)\overline{E_m\left(\frac{2\pi j}{N}\right)} = \frac{1}{N}\sum_{j=0}^{N-1}[\mathrm{e}^{2\pi\mathrm{i}(k-m)/N}]^j$$

如果 N 整除 $k-m$，那么$(k-m)/N$ 是整数并且 $\mathrm{e}^{2\pi\mathrm{i}(k-m)/N}=1$. 因此，每个被加数是 1，它们的平均数也是 1. 另一方面，如果 $k-m$ 不能被 N 整除，那么 $\mathrm{e}^{2\pi\mathrm{i}(k-m)/N}\neq 1$，这时我们可以应用几何级数求和的标准公式：

$$\sum_{j=0}^{N-1}\lambda^j = \frac{\lambda^N-1}{\lambda-1} \qquad (\lambda\neq 1)$$

其结果是

$$\frac{e^{2\pi i(k-m)}-1}{e^{2\pi i(k-m)/N}-1}=0$$ ■

6.12.4 指数多项式

一个次数至多是 n 次的**指数多项式**指的是下列形式的任一函数：

$$P(x)=\sum_{k=0}^{n}c_k e^{ikx}=\sum_{k=0}^{n}c_k E_k(x)=\sum_{k=0}^{n}c_k(e^{ix})^k$$

等式中最后一个表达式说明了这术语的来源，这是因为它表明 P 是变量 e^{ix} 的次数$\leqslant n$ 的多项式. 下面两个结论概括了指数多项式的插值问题.

定理 3(标准正交函数 E_k 定理) 设 E_k 是函数 $E_k(x)=e^{ikx}$，则$\{E_0, E_1, \cdots, E_{N-1}\}$关于

448 (12)式中定义的内积$\langle\cdot, \cdot\rangle_N$ 是标准正交的.

推论 1(指数多项式推论) 在等距结点 $x_j=2\pi j/N$ 上插值给定函数 f 的指数多项式由下列等式给出：

$$P=\sum_{k=0}^{N-1}c_k E_k \quad 其中 \quad c_k=\langle f, E_k\rangle_N \tag{14}$$

证明 用 c_k 的已知公式，我们计算在任意结点 x_v 上指数多项式的值，其结果是

$$\begin{aligned}\sum_{k=0}^{N-1}c_k E_k(x_v) &= \sum_{k=0}^{N-1}\langle f, E_k\rangle_N E_k(x_v)\\ &= \sum_{k=0}^{N-1}\frac{1}{N}\sum_{j=0}^{N-1}f(x_j)\overline{E_k(x_j)}E_k(x_v)\\ &= \sum_{j=0}^{N-1}f(x_j)\frac{1}{N}\sum_{k=0}^{N-1}\overline{E_j(x_k)}E_v(x_k)\\ &= \sum_{j=0}^{N-1}f(x_j)\langle E_v, E_j\rangle_N\\ &= f(x_v)\end{aligned}$$ ■

例 1 利用推论 1，给出 $N=2$ 时插值的显式公式.

解 此时，P 是在结点 0 和 π 上插值 f 的 1 次指数多项式. 由(14)式给出

$$\begin{aligned}P(x) &= \frac{1}{2}[f(0)+f(\pi)]+\frac{1}{2}[f(0)+f(\pi)e^{-i\pi}]e^{ix}\\ &= \frac{1}{2}[f(0)+f(\pi)]+\frac{1}{2}[f(0)-f(\pi)]e^{ix}\end{aligned}$$ ■

推论 2(指数多项式推论) 若 $n<N$，则在有限集

$$x_j=2\pi j/N \qquad (0\leqslant j\leqslant N-1)$$

上得到最小二乘意义下最佳逼近 f 的指数多项式 $\sum_{k=0}^{n}c_k E_k$，其中 $c_k=\langle f, E_k\rangle_N$.

证明 根据定理 3，关于(12)式定义的内积，函数 $x\mapsto e^{ikx}$ 构成一个标准正交系. 于是利用

449 6.8 节中的定理 3 便可完成证明. ■

在推论 1 中，其内容文字的选择暗示所述的指数多项式是唯一的. 为证明这一点，假设

$\sum_{k=0}^{N-1} a_k E_k$ 是在点 $x_0, x_1, \cdots, x_{N-1}$（其中 $x_j = 2\pi j/N$）上插值 f 的指数多项式. 从而

$$\sum_{k=0}^{N-1} a_k E_k(x_j) = f(x_j) \qquad (0 \leqslant j \leqslant N-1)$$

如果给等式两端同时乘以 $E_n(-x_j)$. 然后再对 j 求和，其结果是

$$\sum_{k=0}^{N-1} a_k \sum_{j=0}^{N-1} E_k(x_j) E_n(-x_j) = \sum_{j=0}^{N-1} f(x_j) E_n(-x_j)$$

根据(12)式，可以推得

$$\sum_{k=0}^{N-1} a_k \langle E_k, E_n \rangle_N = \langle f, E_n \rangle_N$$

因为 $\langle E_k, E_n \rangle_N = \delta_{kn}$，所以我们得到

$$a_n = \langle f, E_n \rangle_N = c_n$$

习题 6.12

1. 利用定理 3 和推论 1 中的记号，证明：如果指数多项式 $g(x) = \sum_{k=0}^{N-1} a_k E_k(x)$ 在每个结点 x_j 上取值 0，那么系数 a_k 都为 0.
2. (续)利用上题的结论，用另一种方法证明推论 1 中的插值函数是唯一的.
3. 证明：$E_k E_n = E_{k+n}$ 和 $\overline{E_k} = E_{-k}$.
4. 证明：如果函数 f 和 g 满足
$$f(x_j) = \langle g, E_j \rangle_n \qquad (x_j = 2\pi j/n)$$
那么 $g(x_j) = n\langle f, E_j \rangle_n$.
5. 用一个适当的指数等式的实部和虚部，证明：
$$\frac{1}{n}\sum_{j=0}^{n-1} \cos\frac{2\pi jk}{n} = \begin{cases} 1 & \text{若 } k \text{ 整除 } n \\ 0 & \text{其他} \end{cases}$$
$$\frac{1}{n}\sum_{j=0}^{n-1} \sin\frac{2\pi jk}{n} = 0$$
6. 证明：(12)式中定义的内积满足它后面的三条性质 1，2，3. 并问为什么 $\|\cdot\|_N$ 不是范数？

450

6.13 快速傅里叶变换

傅里叶变换可以把信号分解成为它的组成频率. 类似于一个棱镜把白光分离到它的彩光分频带. 墨镜为减少白光的眩目而只让柔和的绿色光线通过，用同样的方式，傅里叶变换可以用来改变信号而得到满意的效果. 通过分析信号或者系统的分量频率，傅里叶级数和变换在广泛的应用领域找到了它的用途，例如航空器和宇宙飞船制导、数字信号处理、医疗成像、石油和天然气开发、求解微分方程等. 详情可参阅 Briggs and Henson[1995]或者 Walker[1992].

本节致力于三角插值的计算方面. 特别是，为有效地确定 6.12 节等式(14)中的系数，要详尽地阐述称为**快速傅里叶变换**的算法，我们遵循 Stoer and Bulirsch[1980]的阐述.

假设系数 $c_0, c_1, c_2, \cdots, c_{N-1}$ 的定义如同 6.12 节推论 1 中的那样. 我们记

$$c_k = \frac{1}{N}\sum_{j=0}^{N-1} f(x_j)(\lambda^k)^j \qquad (\lambda = e^{-2\pi i/N})$$

从而，c_k 是一个 $N-1$ 次多项式在点 λ^k 处计算的结果；计算出 c_k 大约要用 N 次乘法和 N 次加法的运算成本. 在这个直接的方法中，因为需要计算 N 个系数 c_k，所以构成指数多项式插值的总成本是 $\mathcal{O}(N^2)$次运算.

快速傅里叶变换将把这个计算成本降至一个很合理的数量，那就是 $N\log_2 N$. 下面的表值表明对于一个较大数值的 N 这意味着什么，这样的 N 在信号处理中是常见的：

N	N^2	$N\log_2 N$
1 024	1 048 576	10 240
4 096	16 777 216	49 152
16 384	268 435 456	229 375

定理 1(指数多项式定理) 设 p 和 q 是次数$\leqslant n-1$ 的指数多项式，使得对点 $x_j=\pi j/n$，我们有

$$p(x_{2j}) = f(x_{2j}) \qquad q(x_{2j}) = f(x_{2j+1}) \qquad (0 \leqslant j \leqslant n-1) \tag{1}$$

则在点 x_0，x_1，…，x_{2n-1} 上插值 f 的次数$\leqslant 2n-1$ 的指数多项式由下式给出

$$P(x) = \frac{1}{2}(1+e^{inx})p(x) + \frac{1}{2}(1-e^{inx})q(x-\pi/n) \tag{2}$$

证明 因为 p 和 q 的次数$\leqslant n-1$，而 e^{inx} 是 n 次的，所以 P 的次数$\leqslant 2n-1$. 接下来只需要证明 P 在结点上插值 f 即可. 对 $0\leqslant j\leqslant 2n-1$ 我们有

451

$$P(x_j) = \frac{1}{2}[1+E_n(x_j)]p(x_j) + \frac{1}{2}[1-E_n(x_j)]q(x_j-\pi/n)$$

注意到

$$E_n(x_j) = e^{\pi inj/n} = e^{\pi ij} = \begin{cases} +1 & j \text{ 是偶数} \\ -1 & j \text{ 是奇数} \end{cases}$$

因此对于偶数 j，我们得到 $P(x_j)=p(x_j)=f(x_j)$，而对于奇数 j，我们有

$$P(x_j) = q(x_j-\pi/n) = q(x_{j-1}) = f(x_j)$$

■

定理 2(指数多项式的系数定理) 设定理 1 中给出的多项式的系数为

$$p = \sum_{j=0}^{n-1}\alpha_j E_j \qquad q = \sum_{j=0}^{n-1}\beta_j E_j \qquad P = \sum_{j=0}^{2n-1}\gamma_j E_j$$

则对于 $0\leqslant j\leqslant n-1$，有

$$\gamma_j = \frac{1}{2}\alpha_j + \frac{1}{2}e^{-ij\pi/n}\beta_j \tag{3}$$

$$\gamma_{j+n} = \frac{1}{2}\alpha_j - \frac{1}{2}e^{-ij\pi/n}\beta_j \tag{4}$$

证明 我们将利用(2)式，并且需要 $q(x-\pi/n)$的公式

$$\begin{aligned} q\left(x-\frac{\pi}{n}\right) &= \sum_{j=0}^{n-1}\beta_j E_j\left(x-\frac{\pi}{n}\right) \\ &= \sum_{j=0}^{n-1}\beta_j e^{ij(x-\pi/n)} = \sum_{j=0}^{n-1}\beta_j e^{-i\pi j/n}E_j(x) \end{aligned}$$

这样，根据(2)式，

$$P(x)=\frac{1}{2}[1+E_n(x)]p(x)+\frac{1}{2}[1-E_n(x)]q\left(x-\frac{\pi}{n}\right)$$

$$P=\frac{1}{2}\sum_{j=0}^{n-1}\{(1+E_n)\alpha_jE_j+(1-E_n)\beta_j e^{-i\pi j/n}E_j\}$$

$$=\frac{1}{2}\sum_{j=0}^{n-1}\{(\alpha_j+\beta_j e^{-i\pi j/n})E_j+(\alpha_j-\beta_j e^{-i\pi j/n})E_{n+j}\}$$

于是系数 γ_j 的公式就可以从这个等式获得. ■

例 1 当 $n=1$ 时，利用定理 2 求出 P.

解 (3)式和(4)式给出

$$\gamma_0=\frac{1}{2}(\alpha_0+\beta_0)\quad \gamma_1=\frac{1}{2}(\alpha_0-\beta_0)$$

452

因此，由下式给出 P

$$P=\gamma_0E_0+\gamma_1E_1=\frac{1}{2}(\alpha_0+\beta_0)+\frac{1}{2}(\alpha_0-\beta_0)E_1$$

代入 α_0 和 β_0 的值，我们得到 6.12 节中例 1 的结果：

$$P(x)=\frac{1}{2}[f(0)+f(\pi)]+\frac{1}{2}[f(0)-f(\pi)]e^{ix}$$ ■

6.13.1 分析

为了进一步的分析，设 $R(n)$表示计算点集$\{2\pi j/n: 0\leqslant j\leqslant n-1\}$上插值指数多项式的系数所需最小的乘法运算次数.

定理 3(函数 R 不等式定理) 函数 R 服从不等式

$$R(2^m)\leqslant m2^m$$

证明 我们从建立下列不等式开始讨论

$$R(2n)\leqslant 2R(n)+2n \tag{5}$$

根据定理 2，用 $2n$ 次乘法运算即可从 p 和 q 中的系数得到多项式 P 中所需的系数 γ_j. 实际上，我们需要 n 次乘法用于运算$(1/2)\alpha_j$，$0\leqslant j\leqslant n-1$. 另外 n 次乘法用于计算$(1/2)e^{-ij\pi/n}\beta_j$，$0\leqslant j\leqslant n-1$.(在后面，我们假设已取得了有效的因子$(1/2)e^{-ij\pi/n}$.)因为得到系数 α_0，α_1，…，α_{n-1}需要用 $R(n)$次乘法，并且得到 β_0，β_1，…，β_{n-1}也需要同样的次数，所以可知 P 最多总计需要使用 $2R(n)+2n$ 次乘法运算.

用数学归纳法来证明定理中的不等式. 考虑 $m=0$ 的情况. 这时要用 0 次指数多项式在点 $x_0=0$上插值 f. 它的解是常数 $f(0)$，也不需要做乘法运算. 因而，当 $m=0$ 时定理结论成立. 然后再根据归纳法原理，利用(5)式，计算归纳步骤(从 m 到 $m+1$)如下：

$$R(2^{m+1})=R(2\cdot 2^m)\leqslant 2R(2^m)+2\cdot 2^m$$

$$\leqslant 2(m2^m)+2^{m+1}=(m+1)2^{m+1}$$ ■

作为定理 3 的一个推论，我们看出如果 N 是 2 的方幂，例如 2^m，那么计算指数多项式插值的运算成本服从不等式 $R(N)\leqslant N\log_2 N$. 而用来反复执行定理 1 中程序的算法是快速傅里叶变换.

[453] 定理 1 的内容可以用两个线性算子 L_n 和 T_h 来说明. 对任意的 f，设 L_nf 表示在结点 $2\pi j/n$ $(0\leqslant j\leqslant n-1)$上插值 f 的 $n-1$ 次指数多项式. 设 T_h 是如下定义的**平移算子**：

$$(T_hf)(x)=f(x+h)$$

由 6.12 节推论 1 知

$$L_nf=\sum_{k=0}^{n-1}\langle f,E_k\rangle_n E_k$$

此外，根据定理 1，我们有

$$P=L_{2n}f$$
$$p=L_nf$$
$$q=L_nT_{\pi/n}f$$

根据定理 1 和定理 2，从 L_nf 和 $L_nT_{\pi/n}f$ 可有效地得到 $L_{2n}f$. 当然，这对 $n=1,2,\cdots$都成立.

利用刚才所引入的记号，定理 1 可表示为下面优美的形式

$$L_{2n}f=\frac{1}{2}(1+E_n)L_nf+\frac{1}{2}(1-E_n)T_{-\pi/n}L_nT_{\pi/n}f$$

我们现在的目标是建立快速傅里叶变换算法的一种形式用于计算 L_Nf，其中 $N=2^m$.

定义 1（$P_k^{(n)}$ 的定义） *对给定的函数 f 和指数 $N=2^m$，我们定义*

$$P_k^{(n)}=L_{2^n}T_{2k\pi/N}f\qquad(0\leqslant n\leqslant m,0\leqslant k\leqslant 2^{m-n}-1)\tag{6}$$

如同插值 f 的 2^n-1 次指数多项式那样，$P_k^{(n)}$ 的另一种描述方式如下：

$$P_k^{(n)}\left(\frac{2\pi j}{2^n}\right)=f\left(\frac{2\pi k}{N}+\frac{2\pi j}{2^n}\right)\qquad(0\leqslant j\leqslant 2^n-1)\tag{7}$$

根据习题 6.13.4，与某一个值 k 对应的结点集$(2\pi k/N)+(2\pi j/2^n)$和与另一个值 k 相对应的结点集是不相交的. 直接应用定理 1 表明

$$P_k^{(n+1)}(x)=\frac{1}{2}(1+e^{i2^nx})P_k^{(n)}(x)+\frac{1}{2}(1-e^{i2^nx})P_{k+2^{m-n-1}}^{(n)}\left(x-\frac{\pi}{2^n}\right)\tag{8}$$

用一个**树形图**，我们可以解释这些指数多项式 $P_k^{(n)}$ 是如何相关联的. 假设我们的目标是计算 $P_0^{(3)}$. 按照(8)式，这个函数很容易就从 $P_0^{(2)}$ 和 $P_1^{(2)}$ 获得. 而这两个函数也很容易地依次从 4 个低阶多项式获得，以此类推. 这种联系如图 6-23 所示. [454]

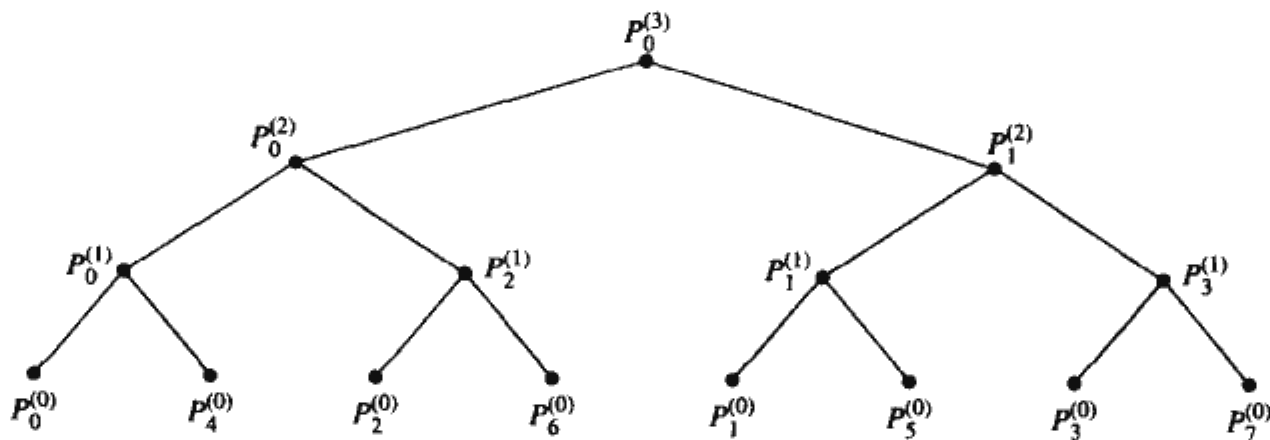


图 6-23 树形图

6.13.2 算法

用 $A_{kj}^{(n)}$ 表示 $P_k^{(n)}$ 的系数. 其中，$0\leqslant n\leqslant m$，$0\leqslant k\leqslant 2^{m-n}-1$，且 $0\leqslant j\leqslant 2^n-1$. 有

$$P_k^{(n)}(x)=\sum_{j=0}^{2^n-1}A_{kj}^{(n)}E_j(x)=\sum_{j=0}^{2^n-1}A_{kj}^{(n)}\mathrm{e}^{\mathrm{i}jx}$$

根据定理 2，下列等式成立

$$A_{kj}^{(n+1)}=\frac{1}{2}\left[A_{kj}^{(n)}+\mathrm{e}^{-\mathrm{i}j\pi/2^n}A_{k+2^{m-n-1},j}^{(n)}\right]$$

$$A_{k,j+2^n}^{(n+1)}=\frac{1}{2}\left[A_{kj}^{(n)}-\mathrm{e}^{-\mathrm{i}j\pi/2^n}A_{k+2^{m-n-1},j}^{(n)}\right]$$

对一固定的 n，因为 $0\leqslant k\leqslant 2^{m-n}-1$ 和 $0\leqslant j\leqslant 2^n-1$，数组 $A^{(n)}$ 需要存储器中 N 个存储单元. 一种完成计算的方法是利用长度为 N 的两个线性数组，一个用来容纳 $A^{(n)}$，另一个容纳 $A^{(n+1)}$. 而在下一步中，一个数组要容纳 $A^{(n+1)}$，而另一个要容纳 $A^{(n+2)}$. 我们称这些数组为 C 和 D. 2 维数组 $A^{(n)}$ 按下列规则储存在 C 中

$$C(2^nk+j)\leftarrow A_{kj}^{(n)}\qquad(0\leqslant k\leqslant 2^{m-n}-1,0\leqslant j\leqslant 2^n-1)$$

同样地，数组 $A^{(n+1)}$ 按下列规则储存在 D 中

$$D(2^{n+1}k+j)\leftarrow A_{kj}^{(n+1)}\qquad(0\leqslant k\leqslant 2^{m-n-1}-1,0\leqslant j\leqslant 2^{n+1}-1)$$

我们一开始就计算并且储存因子 $Z(j)=\mathrm{e}^{-2\pi\mathrm{i}j/N}$. 然后再利用事实 $\mathrm{e}^{-\mathrm{i}j\pi/2^n}=Z(j2^{m-n-1})$. 下面是**快速傅里叶变换算法**：

```
input m
N←2^m
w←e^(-2πi/N)
for k=0 to N−1 do
  Z(k)←w^k
  C(k)←f(2πk/N)
end do
for n=0 to m−1 do
  for k=0 to 2^(m−n−1)−1 do
    for j=0 to 2^n−1 do
        u←C(2^n k+j)
        v←Z(j2^(m−n−1))C(2^n k+2^(m−1)+j)
        D(2^(n+1) k+j)←(u+v)/2
        D(2^(n+1) k+j+2^n)←(u−v)/2
    end do
  end do
  for j=0 to N−1 do
    C(j)←D(j)
  end do
end do
output C(0), C(1), ⋯, C(N)
```

通过详细检查这些伪代码，正如定理 3 给出的那样，我们可以验证有关的乘法运算次数的界为 $m2^m$. 并且在代码的嵌套循环中，注意到 n 呈现值 m；接着 k 呈现值 2^{m-n-1}，而 j 呈现值

455

2^n. 在这部分代码中，只有一条指令与乘法运算有关；也就是计算 v 的指令. 这条指令将会出现很多次，相当于乘积 $m\times 2^{m-n-1}\times 2^n = m2^{m-1}$. 在代码的开始阶段，$Z$ 数组的计算需要 2^m-1 次乘法. 对于任何二进制计算机，乘以 1/2 的运算通常不计在乘法运算次数内，因为它是通过浮点数的指数减 1 来完成的.

例 2 设 $f=\sum_{k=0}^{7}(k+1)E_k$，利用 f 在 8 个等距点上的采样，用快速傅里叶变换重新构造 f.

解 可利用 $m=3$ 时所给定的算法. 一个用来计算 f 值的程序或子程序直接源于已知公式，它运行如下：

```
input x
z←cos x+i sin x
d←8
for k=1 to7 do
  d←dz+8−k
end do
output d
```

当该程序在一台 32-位计算机上运行时，C 数组的初始内容如下(舍入到 5 位数字)：

$$C_0 = 36.000 \qquad C_4 = -4.0000$$
$$C_1 = -4.0000-9.6569\mathrm{i} \qquad C_5 = -4.0000+1.6568\mathrm{i}$$
$$C_2 = -4.0000-4.0000\mathrm{i} \qquad C_6 = -4.0000+4.0000\mathrm{i}$$
$$C_3 = -4.0000-1.6569\mathrm{i} \qquad C_7 = -4.0000+9.6569\mathrm{i}$$

456 C 数组的最终内容是

$$(1.0000, 2.0000, 3.0000, 4.0000, 5.0000, 6.0000, 7.0000, 8.0000)$$ ■

前面的算法只能适用于教学目的. 而作为计算产品的代码还应该进一步地精心改进，例如，用附加的程序设计，我们可以省却 D 数组. 此外，无论因子 $Z(j)$ 是 $+1$ 或者 -1，乘积 $Z(j)C(k)$ 在程序中都应该简单地设计为 $C(k)$ 或者 $-C(k)$. 最后，再编写一个多用途的代码用来妥善处理不是 2 的方幂的那些 N 值.

可编写一些特殊代码用于实变量的实值函数 f，还可编写另外一些代码使其只产生正弦级数或者余弦级数. 有关这个主题的各种分支可参阅本节末的参考文献.

6.13.3 混淆现象和奈奎斯特频率

当我们用样本 $f(x_j)$ 重构函数 f 时，这个过程也有它的局限性问题. 显然，有限样本不可能传递一个可能是无限维线性空间中任一元素的函数所包含的全部信息. 从而，两个不同的连续函数从有限的样本所得到的数据有可能是相同的. 这种现象称为**混淆现象**. 为了研究这一点，我们假设 f 用它的傅里叶级数表示为：

$$f=\sum_{k=-\infty}^{\infty}\langle f,E_k\rangle E_k \tag{9}$$

其中我们用到了下列记号：

$$E_k(x)=\mathrm{e}^{ikx} \qquad \langle f,g\rangle=\frac{1}{2\pi}\int_{-\pi}^{\pi}f(x)\,\overline{g(x)}\,\mathrm{d}x$$

对比之下，次数$\leqslant N-1$的插值指数多项式可表示成

$$P=\sum_{k=0}^{N-1}\langle f,E_k\rangle_N E_k \tag{10}$$

其中的记号是

$$\langle f,g\rangle_N=\frac{1}{N}\sum_{j=0}^{N-1}f(x_j)\,\overline{g(x_j)}\qquad x_j=2\pi j/N$$

比较(9)式和(10)式中的系数，我们给出

$$\langle f,E_m\rangle_N=\left\langle\sum_{k=-\infty}^{\infty}\langle f,E_k\rangle E_k,E_m\right\rangle_N=\sum_{k=-\infty}^{\infty}\langle f,E_k\rangle\langle E_k,E_m\rangle_N$$

根据6.12节定理2，除了$k-m$是N的倍数之外，我们有$\langle E_k,E_m\rangle_N$是0. 这样，如果设$k-m=vN$，其中$v=0$，±1，±2，…，那么我们得到 457

$$\langle f,E_m\rangle_N=\sum_{v=-\infty}^{\infty}\langle f,E_{m+vN}\rangle$$

把这些项按其重要性减少的顺序重写，我们有

$$\langle f,E_m\rangle_N=\langle f,E_m\rangle+\langle f,E_{m+N}\rangle+\langle f,E_{m-N}\rangle+\cdots$$

这表明插值函数(10)中E_m的系数不仅包含我们所需要的那些来自(9)式的项$\langle f,E_m\rangle$，而且还包含那些严格说来属于高次频率的项$\langle f,E_{m\pm vN}\rangle$，$v=1,2,3,\cdots$. 这些项是不需要的而且在重构信号的过程中还会导致失真. 其中第一项是$\langle f,E_N\rangle$，它属于傅里叶级数中的E_N. 但是插值函数(10)式中不会出现E_N项. E_N表示最低频率成分，它只出现在傅里叶级数中但是不在插值函数中. 这种最低频率被称为**奈奎斯特频率**.

6.13.4 计算指数多项式的值

快速傅里叶变换也可以用来计算指数多项式在等距点集上的值. 假设这样的多项式由下列形式给出

$$p(x)=\sum_{j=0}^{n-1}a_jE_j(x)$$

计算p在点

$$t-\frac{2k\pi}{n}\qquad(0\leqslant k\leqslant n-1)$$

上的值. 我们记$x_k=2k\pi/n$，于是

$$\begin{aligned}p(t-x_k)&=\sum_{j=0}^{n-1}a_jE_j(t-x_k)=\sum_{j=0}^{n-1}a_j\mathrm{e}^{\mathrm{i}j(t-x_k)}\\&=\sum_{j=0}^{n-1}a_j\mathrm{e}^{\mathrm{i}jt}\,\overline{E_k(x_j)}=n\langle g,E_k\rangle_n\end{aligned}$$

其中g是一个函数使得

$$g(x_j)=a_j\mathrm{e}^{\mathrm{i}jt}\qquad(0\leqslant j\leqslant n-1)$$

因而，对g应用快速傅里叶变换，得到系数的值$\langle g,E_k\rangle_n$. 当它们被乘以n以后，我们得到$p(t-x_k)$.

快速傅里叶变换的一些信息来源是：Davis and Rabinowitz[1984]、Cooley，Lewis and Welch[1967]、Bloomfield[1976]、Briggs and Henson[1995]、Brigham[1974]、Conte and de Boor[1980]、Kahaner[1970，1978]、Lanczos[1966]、Kahaner，Moler，and Nash[1989]、Elliott and Rao[1982]、Nussbaumer[1982]以及 Scheid[1988].

458

习题 6.13

1. 证明：$T_h E_j$ 是 E_j 的一个数量倍数.
2. 证明：如果 f 和 $f+\lambda E_k$ 是在点 $x_j=2\pi j/N$ 上的被插值函数，其中 k 是 N 的倍数，那么它们是不能相区别的. 因而，为了检验频率$\leqslant\omega$的 f 的所有成分波，所选取的样本频率应该大于 ω. 这是混淆现象另一方面的影响.
3. (**混淆现象**)如果我们只知道函数在离散点上所取的样本，那么两个函数可能会出现相同的结果. 证明：当在 x 的整数值上选取样本时，下面两个函数将会出现相同的结果
$$f(x)=\cos[(n-\alpha)\pi x+\beta] \qquad g(x)=\cos[(n+\alpha)\pi x-\beta]$$
4. 证明：如果 $k\neq r$，那么(7)式中右端的结点集是不相交的.
5. 证明(8)式的正确性.
6. 证明定理 2 的等价形式
$$\langle f,E_j\rangle_{2n}=\frac{1}{2}(u_j+v_j)$$
$$\langle f,E_{j+n}\rangle_{2n}=\frac{1}{2}(u_j-v_j)$$
其中
$$u_j=\langle f,E_j\rangle_n \qquad v_j=e^{-ij\pi/n}\langle T_{\pi/n},E_j\rangle_n$$
这里 $T_{\pi/n}$是用$(T_{\pi/n}f)(x)=f(x+\pi/n)$定义的变换算子.
7. 本题以及后面的三个习题给出了快速傅里叶变换的矩阵公式表示. 固定指标 n 和函数 f. 设 u 和 v 是列向量，其分量如下：
$$u_j=f\left(\frac{2\pi j}{n}\right) \qquad v_j=\langle f,E_j\rangle_n \qquad (0\leqslant j\leqslant n-1)$$
证明：$v=Au$，其中 A 是一个矩阵，具有元素 $A_{jk}=n^{-1}e^{-2\pi ijk/n}(0\leqslant j, k\leqslant n-1)$.
8. (续)证明矩阵 A 只包含 n 个不同的元素. 证明 $n^{1/2}A$ 是一个酉矩阵. 证明 A 是对称矩阵.
9. (续)利用$\mathbb{C}^n$中的欧几里得范数，证明 $\|v\|=n^{-1/2}\|u\|$.
10. (续)沿用以上三题的记号并且记 $A=A^{(n)}$，证明 A 依赖于n. 如果 $u\in\mathbb{C}^{2n}$，令 $u'=(u_0, u_2, \cdots, u_{2n-2})$和 $u''=(u_1, u_3, \cdots, u_{2n-1})$. 证明：对 $0\leqslant k\leqslant n-1$，有
$$(A^{(2n)}u)_k=\frac{1}{2}(A^{(n)}u')_k+\frac{1}{2}e^{-ik\pi/n}(A^{(n)}u'')_k$$
$$(A^{(2n)}u)_{n+k}=\frac{1}{2}(A^{(n)}u')_k-\frac{1}{2}e^{-ik\pi/n}(A^{(n)}u'')_k$$

计算机习题 6.13

编写并测试一个快速傅里叶变换代码，用于计算下列函数的值：
$$p(x)=\sum_{j=0}^{N-1}a_jE_j(x)$$

459

假设 $N=2^m$ 并且系数 a_j 是给定的.

6.14 自适应逼近

自适应逼近的特征是反复剖分函数的定义域，在其较小的子区域上获得更精确的逼近．然后再把多个局部逼近拼接在一起产生分段定义的整体逼近．显然，在自由结点上的样条函数就非常适宜于这种类型的逼近．

6.14.1 一次样条

我们用一个算法来解释上述原理，这个算法通过求一次样条函数来逼近在固定区间$[a,b]$上给定的函数f．假设给定误差容限$\varepsilon>0$，并且我们的目的是求一个一次样条函数S，它满足不等式

$$|f(x)-S(x)|\leqslant\varepsilon,\text{对于 } a\leqslant x\leqslant b \tag{1}$$

样条函数S是分段线性的并且在结点上插值f．在两点α和β上插值f的线性函数由下式给出：

$$\ell(f,\alpha,\beta;x)=\frac{f(\alpha)}{\beta-\alpha}(\beta-x)+\frac{f(\beta)}{\beta-\alpha}(x-\alpha)$$

如果给定一组结点，例如

$$a=t_0<t_1<t_2<\cdots<t_n=b$$

那么f的一次插值样条由下列公式定义

$$S(x)=\begin{cases}\ell(f,t_0,t_1;x) & t_0\leqslant x\leqslant t_1\\ \ell(f,t_1,t_2;x) & t_1\leqslant x\leqslant t_2\\ \vdots & \vdots\\ \ell(f,t_{n-1},t_n;x) & t_{n-1}\leqslant x\leqslant t_n\end{cases}$$

如果S达到不等式(1)中的误差容限，那么我们已经达到目的了．可是，如果

$$\|f-S\|_\infty=\max_{a\leqslant x\leqslant b}|f(x)-S(x)|>\varepsilon$$

那么在区间中的一点或若干点上就超过了误差容限ε．设误差最大的点是ξ，也就是

$$|f(\xi)-S(\xi)|=\|f-S\|_\infty$$

设指标i满足$t_{i-1}\leqslant\xi\leqslant t_i$．在这个区间上，当前的样条函数$S$不符合要求．因而，在区间$[t_{i-1},t_i]$的某些点上的差$|f(x)-\ell(f,t_{i-1},t_i;x)|$超过了$\varepsilon$．因此我们引入新结点$\xi$来剖分这个区间．于是我们用区间$[t_{i-1},\xi]$上的线性函数$\ell(f,t_{i-1},\xi;x)$和区间$[\xi,t_i]$上的线性函数$\ell(f,\xi,t_i;x)$代替$[t_{i-1},t_i]$上的单个线性函数$\ell(f,t_{i-1},t_i;x)$．用这种方法，我们把一个结点集变为另一个结点集，把一个样条逼近变为另一个更好的样条逼近．重复这个过程，一直到样条函数达到误差容限为止． 460

6.14.2 算法

为了编写这个过程的一个算法程序，我们要最大优先地考虑效率．因而在此过程的每一步，我们不仅要储存当前的**结点数组**$t=[t_0,t_1,\cdots,t_n]$，还要储存对应的**纵坐标数组**$[y_0,y_1,\cdots,y_n]$，其中$y_i=f(t_i)$．另外，我们不但要储存**偏差数组**$d=[d_1,d_2,\cdots,d_n]$，其中$d_i=\max\limits_{t_{i-1}\leqslant x\leqslant t_i}|f(x)-\ell(f,t_{i-1},t_i;x)|$，也要储存**极大偏差点数组**$c=[c_1,c_2,\cdots,c_n]$，它

由那些产生极大偏差的点组成. 其中, d_i 是区间$[t_{i-1}, t_i]$内的极大偏差, c_i 是那个区间中产生极大偏差的点. 根据这些储存的量, 我们有

$$\ell(f,t_{i-1},t_i;x)=[y_{i-1}(t_i-x)+y_i(x-t_{i-1})]/(t_i-t_{i-1})$$

$$\|f-S\|_\infty=\max\{d_1,d_2,\cdots,d_n\}$$

$$d_i=|f(c_i)-\ell(f,t_{i-1},t_i;c_i)| \qquad t_{i-1}\leqslant c_i\leqslant t_i$$

在第 n 步, 由于先前的工作, 数组 t, y, d, c 是可利用的. 现在要执行一次检测来判断当前的结点序列是否满足要求. 如果它不满足, 选取指标 i 使得 $d_i=\max\{d_1, d_2, \cdots, d_n\}$. 因为 c_i 是区间$[t_{i-1}, t_i]$内一点, 在这一点上偏差 $|f(x)-S(x)|$ 取极大值, 我们取 c_i 是新结点(前文中记为 ξ). 那么在下一步, $[a, b]$的划分中将呈现两个新区间; 它们是$[t_{i-1}, c_i]$和$[c_i, t_i]$. 储存器中所有指标为 i, $i+1$, $\cdots$, n 的储存数据都向右移位, 而且 t_i 变成 c_i. 换言之, 新结点右边的结点已经重新编号, 与之相应的纵坐标, 偏差, 最大偏差点等也被同样重新编号, 而且把一个新的结点插入到结点数组中, 于是, 包含新结点的区间在新结点处一分为二, 并且要计算这些子区间上相应的新的极大偏差点. 要是有可能的话, 就用这种方式产生一个结点集并且在每个子区间上的极大偏差不超过指定误差容限. 下面是实现这一过程的算法, 用 $t_0=a$ 和 $t_1=b$ 开始:

```
input a, b, ε, M
t0 ← a; t1 ← b; y0 ← f(t0); y1 ← f(t1)
call Max(f, t0, t1, c1, d1)
for n=1 to M-1 do
   选择 i 使得 d_i = max{d_1, d_2, ..., d_n}
   if d_i ≤ ε exit loop
   for j=n to i+1 step -1 do
      t_{j+1} ← t_j
      y_{j+1} ← y_j
      d_{j+1} ← d_j
      c_{j+1} ← c_j
   end do
   t_{i+1} ← t_i
   y_{i+1} ← y_i
   t_i ← c_i
   y_i ← f(c_i)
   call Max(f, t_{i-1}, t_i, c_i, d_i)
   call Max(f, t_i, t_{i+1}, c_{i+1}, d_{i+1})
end do
output n, (t0, t1, ..., tn), (y0, y1, ..., yn), (d1, d2, ..., dn)
```

461

算法仅从两个结点 $t_0=a$ 和 $t_1=b$ 开始. 接下来, 我们确定区间$[t_0, t_1]$上的极大偏差 d_1. 如果 d_1 大于允许误差容限 ε, 则产生极大偏差的点 c_1 变成了一个结点, 结点 t_1 重新编号为 t_2 而 c_1 变成 t_1. 然后算出$[t_0, t_1]$和$[t_1, t_2]$上的极大偏差. 如果其中任何一个超过 ε, 再继续上述过程. 最终, 所有偏差中任何一个都不超过 ε(成功). 或者执行步骤次数超过它的上限 M(失败). 必须要为数组 t, y, d 和 c 中的 M 个分量提供足够的储存容量.

在此算法中，三次出现子程序或过程 Max. Max(f, α, β, c, d)的目的是计算区间[α, β]上 $|f(x)-\ell(f, \alpha, \beta; x)|$ 的极大值，存储极大值于 d 中并且把产生极大值的相应点储存于 c 中. 实际上，c 和 d 的值不需要很高的精度. 一种粗略的方法是选取区间内 11 个点上的 $|f(x)-\ell(f, \alpha, \beta; x)|$ 值，并且采纳其中之一作为所要的点. 下面给出的算法就是做这项工作：

```
procedure Max(f, α, β, c, d)
k←10
h←(β−α)/k
for i=0 to k do
  z_i←f(α+ih)
end do
for i=1 to k−1 do
  z_i←|z_i−(iz_k+(k−i)z_0)/k|
end do
d←0
for i=1 to k−1 do
  if z_i>d then
        d←z_i
        c←α+ih
  end if
end do
return
```

462

在前述的 Max 算法中，首先我们存储 $f(\alpha+ih)$在 z_i 中. 计算出 z_0，z_1，…，z_k 以后，我们有等式

$$\ell(f,\alpha,\beta;x) = \frac{z_0}{kh}(\beta-x)+\frac{z_k}{kh}(x-\alpha)$$

其中 $h=(\beta-\alpha)/k$. 对 $x=\alpha+ih$ 我们有(因为 $\beta=\alpha+kh$)

$$\ell(f,\alpha,\beta;\alpha+ih) = \frac{z_0}{kh}(k-i)h+\frac{z_k}{kh}ih = (iz_k+(k-i)z_0)/k$$

那么存储单元 z_i(算法第 8 行)用来储存

$$|f(\alpha+ih)-\ell(f,\alpha,\beta;\alpha+ih)|$$

例 1 利用区间[0, 1]上函数 $f(x)=\sqrt{x}$以及 $\varepsilon=10^{-2}$检测上面所给出的结合自适应逼近方法的计算机程序. 最终，产生了 10 个结点，密集靠近在 0 点的函数奇异点(具有无穷导数的点). 见图 6-24.

解 10 个结点、f 的值以及极大偏差是：

结点	f 的值	偏差
0.00	0.000	
0.000 729	0.027	0.007
0.002 43	0.049	0.002

0.008 1	0.09	0.003
0.027	0.16	0.005
0.09	0.3	0.01
0.174	0.417	0.005
0.3	0.548	0.004
0.58	0.762	0.009
1.0	1.000	0.008

如果 ε 减小为 10^{-3}，那么这个自适应过程将产生 32 个结点．

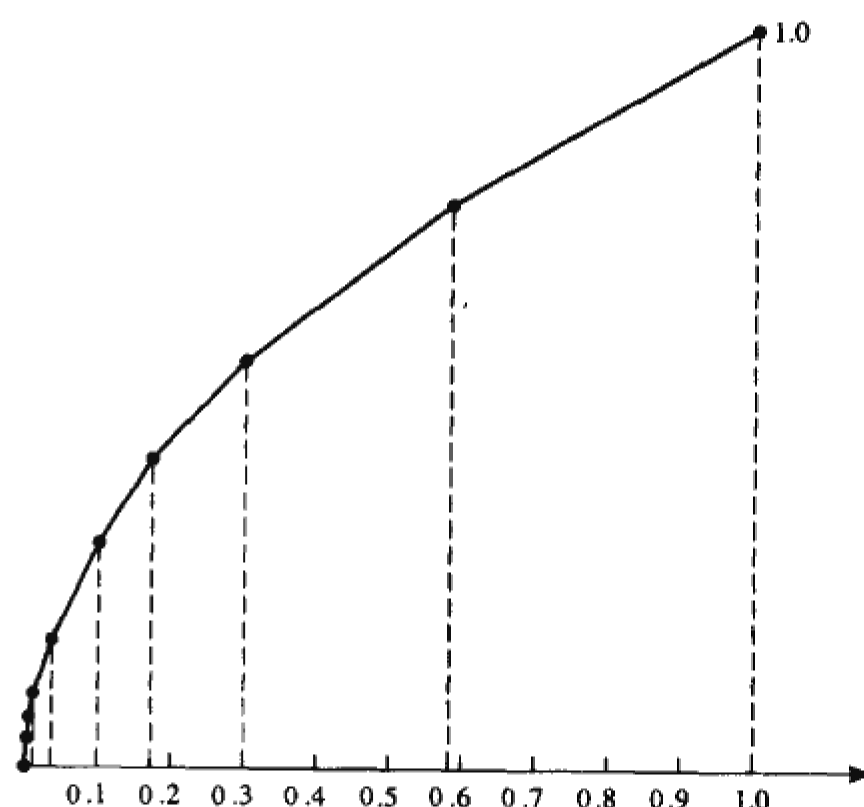


图 6-24 一次自适应样条逼近

■

6.14.3 一般情况

利用一次自适应样条算法所解释的上述原理可抽象化，使得它们能应用于更一般的情况．其过程中的基本要素就是一个局部逼近算子 A．它作用在函数 f 和区间$[\alpha, \beta]$上并且在给定区间上产生 f 的一个逼近．我们用 $A(f, \alpha, \beta; x)$表示这个逼近函数在点 x 的值．在 1 次样条中所使用的运算 $\ell(f, \alpha, \beta; x)$就是这样一个算子．

在自适应过程的每一步，都给定$[a, b]$的一组子区间集，这些区间互不重叠并且覆盖$[a, b]$．若$[\alpha, \beta]$是其中一个子区间并且

$$|f(x)-A(f,\alpha,\beta;x)|\leqslant\varepsilon \quad \text{在区间}[\alpha,\beta]\text{上}$$

则我们称它是满意的．若所有子区间都是满意的，则整体逼近函数 G 就达到了我们的目标，G 是由各个子空间上的函数 $A(f, \alpha, \beta; x)$构成．（注意，并没有保证 G 是连续的．）若子区间$[\alpha, \beta]$不是满意的，则用标准方式剖分该区间．例如，可把中点 $\xi=(1/2)(\alpha+\beta)$添加为结点．另一种方法是选择$[\alpha, \beta]$上误差最大的点 ξ 作为新结点．无论如何，区间$[\alpha, \beta]$会被两个新的

区间$[\alpha, \xi]$和$[\xi, \beta]$所替换．然后再重复这个过程．

习题 6.14

1. 如果对一个单调递增函数使用下面的计算机习题 6.14.1 的过程，试问所需要的子区间个数的下界是什么？答案用 $f(b)-f(a)$和 ϵ 给出．
2. 对函数 $f(x)=\sqrt{x}$，证明：

$$| f(x) - \ell(f,\alpha,\beta;x) |$$

   在点 $x=(1/4)(\sqrt{\beta}+\sqrt{\alpha})^2$ 达到极大值．
3. 如果将自适应算法应用于（完全按课本中所描述的那样）区间$[0, \pi]$上的函数 $f(x)=\sin 10x$，试问会产生什么结果？

计算机习题 6.14

1. 编写和测试一个用于分段常值函数的自适应程序．
2. 编写一个自适应逼近过程的计算机程序，并且把它应用于本节中的例题．给出相应结果的计算机图．
3. 修改自适应算法使其适用于三次样条．
4. 修改过程 Max 使其适用于 y 数组．用其他一些 Max 过程进行实验．

463～464

第7章　数值微分和数值积分

7.1　数值微分和理查森外推

如果给定函数 f 在 $n+1$ 个点 x_0，x_1，…，x_n 上的值，试问可以利用这些信息计算导数 $f'(c)$或者积分 $\int_a^b f(x)\mathrm{d}x$ 的值吗？在某些条件限制下答案是肯定的.

首先观察函数值 $f(x_0)$，$f(x_1)$，…，$f(x_n)$，仅从这些值并不能得到 f 的很多信息，除非还知道 f 属于某种相对较小的函数类. 因而，如果允许函数 f 在所有连续实值函数族中变动，那么这些值 $f(x_i)$几乎是无用的. 在 6 个点上取相同值的三个连续函数如图 7-1 所示.

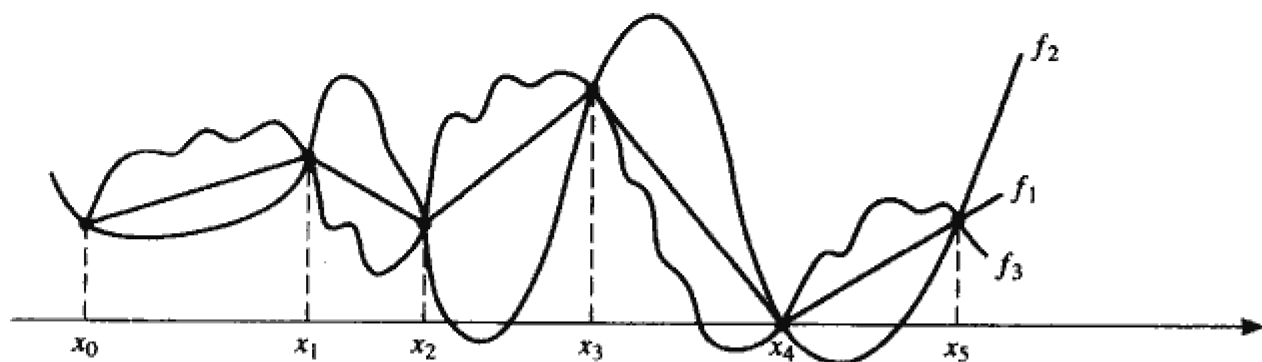


图 7-1　经过 6 点的三个连续函数

另一方面，如果知道 f 是次数至多是 n 次的一个多项式，根据 6.1 节的插值理论，那么 $n+1$ 个点上的函数值可完全确定 f. 这时，可精确地找到 f 而且信心十足地计算 $f'(c)$ 或者 $\int_a^b f(x)\mathrm{d}x$. 然而，在许多实际情况中，已掌握的信息不能完全确定 f，它的导数或积分的数值计算结果也会受到怀疑，除非附带地给出某些相应误差的界.

465

7.1.1　数值微分

我们将通过考察一个数值微分的公式来说明这些问题，该公式直接来自于 $f'(x)$的极限定义：

$$f'(x) \approx \frac{1}{h}[f(x+h) - f(x)] \tag{1}$$

对于线性函数 $f(x)=ax+b$，近似公式(1)是精确的；也就是说对 h 的每个非零值，它给出 $f'(x)$的正确值. 在其他情况下该公式也可能是精确的，但那只是极偶然的情况. 因此，让我们尝试估计这个数值微分公式所包含的误差. 从下列形式的泰勒定理出发：

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(\xi) \tag{2}$$

这里 ξ 是开区间$(x,\ x+h)$内的一点. 为了使等式(2)成立，f 和 f'应该在闭区间$[x,\ x+h]$上连续，而且在对应的开区间上存在 f''. 重新整理等式(2)得到

$$f'(x)=\frac{1}{h}[f(x+h)-f(x)]-\frac{h}{2}f''(\xi) \tag{3}$$

对上述的一大类函数，可利用的误差项与基本数值公式一起出现在上式中，所以等式(3)比等式(1)更有用处. 注意等式(3)中的误差项有两部分：h 的乘幂和 f 的某个高阶导数因子. 后者给出了适合应用误差估计的函数类的表示. 当 h 趋向于 0 时，误差中的 h 项使得整体表达式收敛于 0. 收敛的速度依赖于 h 的乘幂. 这些评注适用于数值分析中的多种误差估计：通常估计式有一项 h 的乘幂和一项因数，它告诉我们，函数必须属于哪种光滑的类型以使得估计有效.

例 1 如果用公式(1)计算函数 $f(x)=\cos x$ 在点 $x=\pi/4$ 的导数，取 $h=0.01$，答案是多少？它的精确度如何？

解 通过计算，我们有

466

$$f'(x)\approx\frac{1}{h}[f(x+h)-f(x)]=\frac{1}{0.01}[0.700\ 000\ 476-0.707\ 106\ 781]$$
$$=-0.710\ 630\ 51$$

等式(3)中的误差项可这样估计：

$$\left|\frac{h}{2}f''(\xi)\right|=0.005\ |\cos\xi|\leqslant 0.005$$

我们还可以得到一个更好的界，取 $\pi/4<\xi<\pi/4+h$，则 $|\cos\xi|<0.707\ 107$. 这给出一个界为 0.003 535 5. 实际误差是

$$-\sin\frac{\pi}{4}+0.710\ 630\ 51=0.003\ 523\ 729$$

■

等式(3)中的项 $-(h/2)f''(\xi)$ 称为**截断误差**. 因为在某阶导数处截断泰勒级数产生了这种误差. 在这种情况下，通过截断级数

$$f(x+h)=f(x)+hf'(x)+\frac{h^2}{2}f''(x)+\frac{h^3}{3!}f'''(x)+\cdots$$

得到近似公式(1)

$$f(x+h)\approx f(x)+hf'(x)$$

正如我们将要看到的那样，在公式(1)以及其他类似公式的应用中，截断误差与舍入误差起着同样重要的作用.

粗看起来，等式(3)表明，为了精确计算 $f'(x)$，步长 h 必须很小，因此我们做一个实验，其中 h 通过给定的一个值的序列收敛于 0，计算出相应的 $f'(x)$ 的近似值. 取 $f'(x)=\tan^{-1}x$ 以及点 $x=\sqrt{2}$. 结果应该是 $f'(x)=(x^2+1)^{-1}$ 在点 $\sqrt{2}$ 处的值，为 1/3. 下面是它的一个算法：

```
f(x): = tan⁻¹x
input s←√2; h←1; M←26
F₁←f(s)
for k = 0 to M do
    F₂←f(s+h)
    d←F₂ − F₁
    r←d/h
    output k, h, F₂, F₁, d, r
```

```
    h←h/2
end do
```

32 位计算机的一些输出信息如下：

k	h	F_2	F_1	d	r
4	0.62×10^{-1}	0.975 550 95	0.955 316 60	0.020 234 35	0.323 749 54
12	0.24×10^{-3}	0.955 397 96	0.955 316 60	0.000 081 36	0.333 251 95
20	0.95×10^{-6}	0.955 316 90	0.955 316 60	0.000 000 30	0.312 500 00
24	0.60×10^{-7}	0.955 316 66	0.955 316 60	0.000 000 06	1.000 000 00
26	0.15×10^{-7}	0.955 316 60	0.955 316 60	0.000 000 00	0.000 000 00

在每一行中，d 是差值 F_2-F_1，r 是比值 d/h. 由于**减法相消**，d 的有效数字逐渐减少，直到最后 $d=0$ 和 $r=0$ 为止. 当 $k=12$ 时得到 r 的最佳值，如果四舍五入保留小数后四位，则它有 4 位正确数字. 在该 k 值处，我们注意到 d 有 4 位有效数字. 随着 k 的增加，d 中有效数字的个数在减少. 当然，r 的有效数字不会比 d 的更多. 因而，当 h 很小时，舍入误差阻挠我们得到精确值. 为了使 d 有更好的精度，需要 F_1 和 F_2 有更高的精度，因此在基本计算中就需要更高的精度. 可以使用多精度或者在一台长字长的计算机上执行计算操作. 有时答案从某一步开始变坏，这个步数与机器中使用的字长(更确切地说是单位舍入误差)有关. [467]

数值微分公式最重要的应用是在微分方程的数值解中. 常见的策略是用近似代替导数，正如公式(1)中给出的那样. 这种数值微分公式的精度经常简单地用误差项中 h 的乘幂来判断. 因为 h 总是一个很小的数，所以 h 的乘幂次数越高精度越好. 用这种估计方法，由于公式(2)中的误差是 $\mathcal{O}(h)$，所以其结果是不好的. 一个较好的公式是

$$f'(x) \approx \frac{1}{2h}[f(x+h)-f(x-h)] \tag{4}$$

它由泰勒定理的两种情况导出：

$$f(x+h) = f(x)+hf'(x)+\frac{h^2}{2}f''(x)+\frac{h^3}{3!}f'''(\xi_1) \tag{5}$$

$$f(x-h) = f(x)-hf'(x)+\frac{h^2}{2}f''(x)-\frac{h^3}{3!}f'''(\xi_2) \tag{6}$$

两式相减并且重新整理，得到

$$f'(x) = \frac{1}{2h}[f(x+h)-f(x-h)]-\frac{h^2}{12}[f'''(\xi_1)+f'''(\xi_2)] \tag{7}$$

由于误差中含有 h^2 项，所以这是一个更受欢迎的结果. 可是，注意到误差中出现了 f'''. 所以当 f''' 存在时，该误差项是可应用的.

我们增加一个小小的假定：如果函数 f''' 在 $[x-h,\ x+h]$ 上连续，那么可以简化(7)式中的误差项. 设 M 和 m 分别表示 f''' 在区间 $[x-h,\ x+h]$ 上的最大值和最小值，那么 $f'''(\xi_1)$，$f'''(\xi_2)$ 及 $c\equiv[f'''(\xi_1)+f'''(\xi_2)]/2$ 都在区间 $[m,\ M]$ 中. 因为 f''' 是连续的，假定在 $[x-h,\ x+h]$ 内某点 ξ 上取值是 c，因此，

$$f'''(\xi) = \frac{1}{2}[f'''(\xi_1)+f'''(\xi_2)]$$

把上式代入(7)式，其结果是

468

$$f'(x)=\frac{1}{2h}[f(x+h)-f(x-h)]-\frac{h^2}{6}f'''(\xi) \tag{8}$$

给等式(5)和(6)增添一项，然后两式相加可得到二阶导数的一个重要公式. 重新整理并且应用前面的方法，我们有

$$f''(x)=\frac{1}{h^2}[f(x+h)-2f(x)+f(x-h)]-\frac{h^2}{12}f^{(4)}(\xi) \tag{9}$$

其中 $\xi\in(x-h,\ x+h)$. 该公式常用在二阶微分方程的数值解中.

例 2 用计算机近似计算 $f'(x)$，其中 $f(x)=\tan^{-1}x$ 及 $x=\sqrt{2}$. 用(8)式并且步长 h 趋向于 0. 要记得正确值是 1/3.

解 一个适宜的算法如下

```
f(x): = tan^-1 x
input s←√2; h←1; M←26
for k = 0 to M do
    F2←f(s+h)
    F1←f(s−h)
    d←F2−F1
    r←d/(2h)
    output k, h, F2, F1, d, r
    h←h/2
end do
```

32 位计算机的一些输出信息如下：

k	h	F_2	F_1	d	r
2	0.25	1.029 726 74	0.861 129 82	0.168 596 92	0.337 193 85
10	$0.976\,5\times10^{-3}$	0.955 641 99	0.954 990 92	0.000 651 06	0.333 343 51
18	$0.381\,5\times10^{-5}$	0.955 317 86	0.955 315 35	0.000 002 50	0.328 125 00
26	$0.149\,0\times10^{-7}$	0.955 316 60	0.955 316 60	0.000 000 00	0.000 000 00

我们再一次看到，由于**减法相消**，当 h 趋向于 0 时精度明显地下降. 在 $k=9$ 时输出信息开始显示出这一现象. F_1 和 F_2 的值非常接近，因此它们的差值 d 的有效数字严重丢失. 最后，当 $h\to0^+$ 时，在机器中 F_1 和 F_2 的值将会相等，导数的值也将是 0. 在 $k=26$ 时出现这种情况. 在不同字长的计算机上，这种情况会在不同的 k 值处出现. ■

因为在程序进程中数据的误差会扩大，所以只凭经验获得函数的数值微分是一个冒险的过程. 这一点很容易从(8)式中看出. 例如，如果样本点 $x\pm h$ 是精确确定的而纵坐标 $f(x\pm h)$是非精确的，那么纵坐标上的误差将被乘以 $1/(2h)$倍. 因为 h 很小，所以误差的影响会很大.（这一现象不会在数值积分中出现.）因此，要避免或者格外谨慎地使用来自经验数据的数值微分.

469

7.1.2 通过多项式插值的微分

数值微分和积分的一般方法可基于多项式插值. 假设有函数 f 在点 x_0，x_1，…，x_n 上的 $n+1$ 个函数值. 结点 x_i 上 f 的插值多项式可写为 6.1 节中的拉格朗日型(9)式. 我们把那一节

中定理 2 给出的误差项也算在内，得到

$$f(x)=\sum_{i=0}^{n}f(x_i)\ell_i(x)+\frac{1}{(n+1)!}f^{(n+1)}(\xi_x)w(x) \tag{10}$$

这里我们已有 $w(x)=\prod_{i=0}^{n}(x-x_i)$. 对(10) 式求导数，有

$$f'(x)=\sum_{i=0}^{n}f(x_i)\ell'_i(x)+\frac{1}{(n+1)!}f^{(n+1)}(\xi_x)w'(x)+\frac{1}{(n+1)!}w(x)\frac{\mathrm{d}}{\mathrm{d}x}f^{(n+1)}(\xi_x)$$

如果 x 是一个结点，例如 $x=x_\alpha$，由于 $w(x_\alpha)=0$，那么化简上述等式，其结果是

$$f'(x_\alpha)=\sum_{i=0}^{n}f(x_i)\ell'_i(x_\alpha)+\frac{1}{(n+1)!}f^{(n+1)}(\xi_{x_\alpha})w'(x_\alpha)$$

通过计算 $w'(x_\alpha)$该等式还可以化简. 为此，我们注意到

$$w'(x)=\sum_{i=0}^{n}\prod_{\substack{j=0\\ j\neq i}}^{n}(x-x_j),\quad \text{所以 } w'(x_\alpha)=\prod_{\substack{j=0\\ j\neq \alpha}}^{n}(x_\alpha-x_j)$$

最后带有误差项的微分公式是

$$f'(x_\alpha)=\sum_{i=0}^{n}f(x_i)\ell'_i(x_\alpha)+\frac{1}{(n+1)!}f^{(n+1)}(\xi_{x_\alpha})\prod_{\substack{j=0\\ j\neq \alpha}}^{n}(x_\alpha-x_j) \tag{11}$$

这个公式特别适宜于非等距结点.

例 3 当 $n=2$ 及 $\alpha=1$ 时，给出(11)式的明确形式.

解 此时，拉格朗日插值的 3 个基函数是

$$\ell_0(x)=\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}$$

$$\ell_1(x)=\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}$$

$$\ell_2(x)=\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}$$

470

它们的导数是

$$\ell'_0(x)=\frac{2x-x_1-x_2}{(x_0-x_1)(x_0-x_2)}$$

$$\ell'_1(x)=\frac{2x-x_0-x_2}{(x_1-x_0)(x_1-x_2)}$$

$$\ell'_2(x)=\frac{2x-x_0-x_1}{(x_2-x_0)(x_2-x_1)}$$

在 x_1 上赋值，我们得到

$$\ell'_0(x_1)=\frac{x_1-x_2}{(x_0-x_1)(x_0-x_2)}$$

$$\ell'_1(x_1)=\frac{2x_1-x_0-x_3}{(x_1-x_0)(x_1-x_2)}$$

$$\ell'_2(x_1)=\frac{x_1-x_0}{(x_2-x_0)(x_2-x_1)}$$

因而带有误差项的数值微分公式是

$$\begin{aligned} f'(x_1) &= f(x_0)\frac{x_1-x_2}{(x_0-x_1)(x_0-x_2)} \\ &+ f(x_1)\frac{2x_1-x_0-x_2}{(x_1-x_0)(x_1-x_2)} \\ &+ f(x_2)\frac{x_1-x_0}{(x_2-x_0)(x_2-x_1)} \\ &+ \frac{1}{6}f'''(\xi_{x_1})(x_1-x_0)(x_1-x_2) \end{aligned} \tag{12}$$

■

例 4 在例3中，如果结点是等距的，公式的结果是什么？

解 设 $x_0=x_1-h$ 及 $x_2=x_1+h$. 则由(12)式知(取 $x=x_1$)

$$f'(x) = f(x-h)\left(\frac{-1}{2h}\right) + f(x+h)\left(\frac{1}{2h}\right) - \frac{1}{6}f'''(\xi_x)h^2$$

这与前面推导出的(8)式相同. ■

7.1.3 理查森外推

我们现在引入一个称为**理查森外推**的过程，并介绍如何利用它来巧妙地改进数值公式的精度. 把(5)式和(6)式扩展到具有高阶项. 假设 $f(x)$用它的泰勒级数表示为

$$f(x+h) = \sum_{k=0}^{\infty}\frac{1}{k!}h^k f^{(k)}(x) \tag{13}$$

[471]

$$f(x-h) = \sum_{k=0}^{\infty}\frac{1}{k!}(-1)^k h^k f^{(k)}(x) \tag{14}$$

如果第一个等式减去第二个等式，则消去了所有 k 是偶数的项，得

$$f(x+h)-f(x-h) = 2hf'(x) + \frac{2}{3!}h^3 f'''(x) + \frac{2}{5!}h^5 f^{(5)}(x) + \cdots$$

重新整理得

$$\begin{aligned} f'(x) &= \frac{1}{2h}[f(x+h)-f(x-h)] \\ &\quad - \left[\frac{1}{3!}h^2 f^{(3)}(x) + \frac{1}{5!}h^4 f^{(5)}(x) + \frac{1}{7!}h^6 f^{(7)}(x) + \cdots\right] \end{aligned}$$

这个等式具有形式

$$L = \varphi(h) + a_2h^2 + a_4h^4 + a_6h^6 + \cdots \tag{15}$$

其中 L 表示 $f'(x)$，$\varphi(h)$表示数值微分公式(4)；即

$$\varphi(h) = \frac{1}{2h}[f(x+h)-f(x-h)]$$

其中 x 是指定的数值，例如 s，我们要计算的就是这一点上的导数. 下面设计的数值过程用于估计 L. 对于 $h>0$，可计算函数 φ 的值，但不能计算 $h=0$ 时 φ 的值. 因而，在计算 L 的过程中，我们只能令 h 趋向于0. 对每个 $h>0$，级数的项 $a_2h^2+a_4h^4+\cdots$给出了误差. 假设 $a_2\neq 0$，可以看出当 h 充分小时，第一项 a_2h^2 大于其他项. 因此要设法消去这一占优项 a_2h^2. 我们的

分析仅仅是建立在(15)式的基础上，并且它可应用于其他数值过程；特别是它可用于 7.4 节中的梯形法则.

用 $h/2$ 替换(15)式中的 h 得到

$$L = \varphi(h/2) + a_2h^2/4 + a_4h^4/16 + a_6h^6/64 + \cdots \tag{16}$$

(15)式减去 4 倍的(16)式，可消去误差级数中的第一项 a_2h^2. 结果如下：

$$\begin{array}{l} L = \varphi(h) + a_2h^2 + a_4h^4 + a_6h^6 + \cdots \\ \underline{4L = 4\varphi(h/2) + a_2h^2 + a_4h^4/4 + a_6h^6/16 + \cdots} \\ 3L = 4\varphi(h/2) - \varphi(h) - 3a_4h^4/4 - 15a_6h^6/16 - \cdots \end{array}$$

因此我们有

$$L = \frac{4}{3}\varphi(h/2) - \frac{1}{3}\varphi(h) - a_4h^4/4 - 5a_6h^6/16 - \cdots \tag{17}$$

(17)式表达了理查森外推的第一步. 它表明 $\varphi(h)$ 和 $\varphi(h/2)$ 的一个简单组合提供了一个计算 L 的方法，它具有精度 $\mathcal{O}(h^4)$.

472

例 5 结合理查森外推，重新计算例 2 中的导数.

解 一个适宜的算法如下：

```
f(x): = tan^-1(x)
input s←√2; h←1; M←30
for k = 0 to M do
    d_k←[f(s+h) − f(s−h)]/(2h)
    h←h/2
end do
for k = 1 to M do
    r_k←d_k + (d_k − d_{k−1})/3
end do
output [k, d_k, r_k : 0≤k≤M]
```

下面给出几行输出信息：

k	d_k	r_k
2	0.337 193 85	0.333 334 80
4	0.333 574 77	0.333 333 64
8	0.333 328 25	0.333 325 71
16	0.332 031 25	0.331 380 22
26	0.000 000 00	0.000 000 00

与例 2 中相应输出信息比较，这里得到两位数以上的精度.

刚才给出的例题是人为的，因为我们已经知道正确答案，并用它去确定第三列中的哪个值最精确. 在实际情况中，要考虑计算 d_k 时所丢失的有效数字的数目. 显然，r_k 不可能比 d_k 有更多的有效数字. 凭心而论，并不能保证当 k 无限增大时 r_k 会变得更加精确. ■

对读者来说可能会出现这样的情况，对(15)式所完成的过程现在可以用于(17)式(做适当的修改). 相应的做法如下：在(17)式中令

$$\psi(h) = \frac{4}{3}\varphi(h/2) - \frac{1}{3}\varphi(h)$$

则

$$L = \psi(h) + b_4h^4 + b_6h^6 + \cdots$$
$$L = \psi(h/2) + b_4h^4/16 + b_6h^6/64 + \cdots$$

此时

$$\begin{aligned} L &= \psi(h) + b_4h^4 + b_6h^6 + \cdots \\ 16L &= 16\psi(h/2) + b_4h^4 + b_6h^6/4 + \cdots \\ \hline 15L &= 16\psi(h/2) - \psi(h) - 3b_6h^6/4 - \cdots \end{aligned}$$

473

因而，我们有

$$L = \frac{16}{15}\psi(h/2) - \frac{1}{15}\psi(h) - b_6h^6/20 - \cdots \tag{18}$$

再一次重复这个过程，在(18)式中令

$$\theta(h) = \frac{16}{15}\psi(h/2) - \frac{1}{15}\psi(h)$$

使得

$$L = \theta(h) + c_6h^6 + c_8h^8 + \cdots$$

用上述相同的方法可得

$$L = \frac{64}{63}\theta(h/2) - \frac{1}{63}\theta(h) - 3c_8h^8/252 - \cdots$$

事实上，可执行任意多步来得到不断增加精确度的公式. 下面是完整的算法，即允许执行 M 步的**理查森外推算法**：

1. 选取一个方便的 h 值(例如 $h=1$)并且计算 $M+1$ 个数

$$D(n,0) = \varphi(h/2^n) \quad (0 \leqslant n \leqslant M)$$

2. 用下列公式计算

$$D(n,k) = \frac{4^k}{4^k-1}D(n,k-1) - \frac{1}{4^k-1}D(n-1,k-1) \tag{19}$$

这里 $k=1, 2, \cdots, M$，$n=k, k+1, \cdots, M$.

注意到 $D(0, 0)=\varphi(h)$，$D(1, 0)=\varphi(h/2)$ 和 $D(1, 1)=\psi(h)$，用 $h/2$ 重复替换 h，则 $D(n, 1)$ 与(17)式一致. 类似地，$D(n, 2)$ 与(18)式一致，依此类推，根据我们的假设和计算，显然有

$$D(n,0) = L + \mathcal{O}(h^2)$$
$$D(n,1) = L + \mathcal{O}(h^4)$$
$$D(n,2) = L + \mathcal{O}(h^6)$$
$$D(n,3) = L + \mathcal{O}(h^8)$$

下面的定理中证明的一般结果是

$$D(n,k-1) = L + \mathcal{O}(h^{2k}) \qquad \text{因为 } h \to 0$$

在这方面，我们再次强调上述分析可以应用于任何满足(15)式的数值过程. 位于(15)式前面的

等式只是一个特例，仅用来说明在一个特殊的数值过程中它是怎样出现的.

如果(15)式成立，那么理查森算法是有效的，从这种意义上说 $D(n, k)$阵列中的相继列将显示出更高阶的收敛性. 这就是下述定理的内容.

474

定理 1(理查森外推定理) 理查森外推算法中定义的 $D(n, k)$服从下列形式的等式：

$$D(n,k-1) = L + \sum_{j=k}^{\infty} A_{jk}(h/2^n)^{2j} \tag{20}$$

证明 当 $k=1$ 时，利用 $D(n, 0)$的定义及(15)式验证(20)式成立：

$$D(n,0) = \varphi(h/2^n) = L - \sum_{j=1}^{\infty} a_{2j}(h/2^n)^{2j}$$

因而，可设 $A_{j1}=-a_{2j}$，现在对 k 作数学归纳法. 假设对 $k-1$(20)式成立，在此基础上证明它对 k 也成立. 根据(19)式和(20)式，有

$$D(n,k) = \frac{4^k}{4^k-1}\left[L + \sum_{j=k}^{\infty} A_{j,k}\left(\frac{h}{2^n}\right)^{2j}\right] - \frac{1}{4^k-1}\left[L + \sum_{j=k}^{\infty} A_{j,k}\left(\frac{h}{2^{n-1}}\right)^{2j}\right]$$

它可化简为

$$D(n,k) = L + \sum_{j=k}^{\infty} A_{j,k}\left[\frac{4^k-4^j}{4^k-1}\right]\left(\frac{h}{2^n}\right)^{2j} \tag{21}$$

从而，$A_{j,k+1}$应该定义为

$$A_{j,k+1} = A_{j,k}\left[\frac{4^k-4^j}{4^k-1}\right]$$

注意到 $A_{k,k+1}=0$，因而(21)式可以写为

$$D(n,k) = L + \sum_{j=k+1}^{\infty} A_{j,k+1}(h/2^n)^{2j}$$ ■

根据 $D(n, 0)$和 $D(n, k)$的已知公式，我们构造三角形阵列：

$$\begin{matrix} D(0,0) \\ D(1,0) & D(1,1) \\ D(2,0) & D(2,1) & D(2,2) \\ \vdots & \vdots & \vdots & \ddots \\ D(M,0) & D(M,1) & D(M,2) & \cdots & D(M,M) \end{matrix}$$

产生该三角形阵列的算法在下面给出，在该算法中涉及一个函数 φ，该函数一定要单独编码(即利用子程序或程序).

475

```
input h, M
for n = 0 to M do
    D(n, 0) ← φ(h/2^n)
end do
for k = 1 to M do
    for n = k to M do
        D(n, k) ← D(n, k−1) + [D(n, k−1) − D(n−1, k−1)]/(4^k − 1)
    end do
end do
output D(n, k)  (0≤n≤M, 0≤k≤n)
```

例 6 利用刚才所述的理查森外推算法，重新计算例 5 中的导数.

解 利用例 2，这个算法的某些输出信息如下：

n	$D(n,0)$	$D(n,1)$	$D(n,2)$	$D(n,3)$	$D(n,4)$
0	0.392 699 1				
1	0.348 771 0	0.334 128 3			
2	0.337 193 8	0.333 334 8	0.333 281 9		
3	0.334 298 1	0.333 332 9	0.333 332 8	0.333 333 6	
4	0.333 574 8	0.333 333 6	0.333 333 7	0.333 333 7	0.333 333 7

该算法近似地得到了与例 5 中相同的精度. 由于减法相消性，最终也产生了无意义的结果. ■

近几年，已经开发出了一些**自动微分**的软件工具，Bischof，Carle，Khademi，and Mauer [1994]编写了一个名为 ADIFOR 的软件系统，把链式法则应用于初等指令来计算导数. 他们的目的之一是几乎不需要人为影响来产生有效导数的编码，与软件的黑盒子部件一致. 有关通过计算机程序计算函数导数的技术，参见 Griewank and Corliss[1991].

习题 7.1

1. 详细完成(11)式的求导过程.

2. 设 f 是一个 n 次多项式. 如果已知 f 在 n 个点上的值，试问肯定可以估计 $f'(c)$或者 $\int_a^b f(x)\mathrm{d}x$ 吗？你的答案将与 c，a 和 b 有关.

3. 设一个数值过程为

$$L = \varphi(h) + \sum_{j=1}^{\infty} a_j h^j$$

解释此时理查森外推是如何工作的，在这种情况下，对理查森外推证明一个与本节定理 1 类似的结果.

476

4. 对下列情况给出(11)式的显式形式：

a. $n=0$；$\alpha=0$

b. $n=1$；$\alpha=0$ 及 $\alpha=1$

c. $n=1$；$\alpha=0$ 及 $\alpha=2$

5. $f''(x)$ 的公式(9) 经常用于微分方程的数值解中. 根据 $f(x+h)$ 和 $f(x-h)$ 的泰勒级数，证明该公式中的误差具有形式 $\sum_{n=1}^{\infty} a_{2n}h^{2n}$. 明确地求出系数 a_{2n}. 同时，推导出(9) 式中的误差项.

6. 推导出下列近似导数的两个公式，并且通过建立其误差项证明两者都是 $\mathcal{O}(h^4)$.

$$f'(x) \approx \frac{1}{12h}[-f(x+2h)+8f(x+h)-8f(x-h)+f(x-2h)]$$

$$f''(x) \approx \frac{1}{12h^2}[-f(x+2h)+16f(x+h)-30f(x)+16f(x-h)-f(x-2h)]$$

7. 推导出下列近似三阶导数的两个公式. 求出它们的误差项. 试问哪一个公式更精确？

$$f'''(x) \approx \frac{1}{h^3}[f(x+3h)-3f(x+2h)+3f(x+h)-f(x)]$$

$$f'''(x) \approx \frac{1}{2h^3}[f(x+2h)-2f(x+h)+2f(x-h)-f(x-2h)]$$

8. (续)对下列四阶导数公式完成上题中的要求：

$$f^{(4)} \approx \frac{1}{h^4}[f(x+4h)-4f(x+3h)+6f(x+2h)-4f(x+h)+f(x)]$$

$$f^{(4)} \approx \frac{1}{h^4}[f(x+2h)-4f(x+h)+6f(x)-4f(x-h)+f(x-2h)]$$

9. 证明：在理查森外推中，

$$D(2,2)=\frac{16}{15}\psi(h/2)-\frac{1}{15}\psi(h)$$

10. 如果

$$L=x_n+a_1n^{-1}+a_2n^{-2}+a_3n^{-3}+\cdots$$

说明如何利用使用 x_n 和 x_{n^2} 的理查森外推.

11. 证明或者否定：

a. 若 $L-x_n=\mathcal{O}(n^{-1})$，则 $L-(2x_{2n}-x_n)=\mathcal{O}(n^{-2})$.

b. 若 $L-x_n=\mathcal{O}(n^{-1})$，则 $L-x_{n^2}=\mathcal{O}(n^{-2})$.

讨论这个问题的数值结果.

12. 如果

$$L=\varphi(h)+a_1h+a_3h^3+a_5h^5+\cdots$$

说明如何利用理查森外推.

13. 假设 $L=\lim_{h\to 0}f(h)$ 以及 $L-f(h)=c_6h^6+c_9h^9+\cdots$. 试问 $f(h)$ 和 $f(h/2)$ 的什么样的组合是 L 的最佳估计?

14. 利用泰勒级数，推导出下列近似的误差项

$$f'(x)\approx\frac{1}{2h}[-3f(x)+4f(x+h)-f(x+2h)]$$

477

15. 应用理查森外推于

$$f'(x)=\frac{1}{2h}[f(x+h)-f(x-h)]-\frac{h^2}{6}f'''(x)-\frac{h^4}{120}f^{(5)}(x)-\cdots$$

推导出一个 $\mathcal{O}(h^4)$ 阶的数值微分公式. 给出 $\mathcal{O}(h^4)$ 阶的误差项.

16. 利用泰勒级数展开式，推导出下列公式的误差项

$$f''(x)\approx\frac{1}{h^2}[f(x)-2f(x+h)+f(x+2h)]$$

17. 建立下列形式的一个公式

$$f''_n\approx\frac{1}{h^2}[Af_{n+3}+Bf_{n+2}+Cf_{n+1}+Df_n]$$

其中，$f_{n+i}=f(x_n+ih)$.

18. 推导出近似表达式

$$f'(x_n)\approx\frac{3f(x_n)-4f(x_{n-1})+f(x_{n-2})}{3x_n-4x_{n-1}+x_{n-2}}$$

并且证明当 $h\to 0$ 时误差项是 $\mathcal{O}(h^2)$. 其中，$f_{n+i}=f(x_n+ih)$.

计算机习题 7.1

1. 对课本中重复地使用理查森外推计算 $f'(x)$ 值的算法，编写出它的程序. 对下列情况测试你的程序：

a. $\ln x$，在点 $x=3$.

b. $\tan x$，在点 $x=\sin^{-1}(0.8)$.

c. $\sin\left(x^2+\frac{1}{3}x\right)$，在点 $x=0$.

2. 利用公式(9)以及反复的理查森外推，编写并测试用于计算 $f''(x)$的一个程序.
3. 在公式(1)的一种典型用法中，舍入误差(主要由于减法相消性)将表现为 αh^{-1}，举例检验这个论断(并且估计 α 的值). 可能需要使用多精度的计算机程序模拟这个计算.
4. 利用三次样条，推导并测试近似计算 f'和 f''的一个算法.
5. 利用习题 7.1.18 中的近似表达式，给出一个类似于正割法的迭代公式，对若干函数检验该方法的数值性能，并且把它与正割法做比较.

7.2 基于插值的数值积分

数值积分是一个产生某集合上一个函数的积分数值的过程. 下面是一些积分的例子，它们 478 可以用适当的计算机例行程序计算：

$$\int_0^2 e^{-x^2}\,dx$$

$$\int_0^1\int_0^1 \sin(xye^x)\,dx\,dy$$

$$\int_0^1\int_{x^2}^x \tan(xy^2)\,dy\,dx$$

$$\int_0^\pi \cos(3\cos\theta)\,d\theta$$

这些积分问题不可按照初等微积分学习的技巧处理. 那些技巧依赖于**反微分**. 因而，为了用微积分得到下列积分的值

$$\int_a^b f(x)\,dx$$

我们首先产生一个具有性质 $F'=f$ 的函数 F. 进而有

$$\int_a^b f(x)\,dx = F(b) - F(a)$$

例如，有

$$\int_1^4 x^2\,dx = \frac{1}{3}x^3\Big|_1^4 = \frac{1}{3}4^3 - \frac{1}{3} = 21$$

由 $F(x)=(1/3)x^3$ 给出的函数 F 是函数 $f(x)=x^2$ 的反导数. 很多初等函数没有简单的反导数. 一个很好的例子是 $f(x)=e^{x^2}$. 该函数的一个反导数是

$$F(x) = \sum_{k=0}^{\infty} \frac{x^{2k+1}}{(2k+1)k!}$$

(这个等式的出处见 6.7 节中(3)式.)

数值计算积分

$$\int_a^b f(x)\,dx \tag{1}$$

的一个有效策略是用另一个函数 g 替换 f，其中 g 与 f 非常近似并且容易积分. 简单地说，由 $f\approx g$ 得到

479

$$\int_a^b f(x)\,dx \approx \int_a^b g(x)\,dx$$

读者此刻会想到多项式是函数 g 的很好的选择．的确，g 可以是某个结点集上插值 f 的多项式．当然，f 的多项式逼近也可由其他方法得到，例如，用截断泰勒级数的方法．其次，读者可以回想 6.7 节中类似的做法．例如，我们有

$$\int_0^1 e^{x^2}\,dx \approx \int_0^1 \sum_{k=0}^{n} \frac{x^{2k}}{k!}\,dx = \sum_{k=0}^{n} \frac{1}{(2k+1)k!}$$

无论如何，仅需要被积函数求值的一般过程是我们所想往的．基于插值的方法可以实现这个愿望．样条函数也可用来插值 f，而且很容易计算样条函数的积分．

7.2.1 通过多项式插值的积分

现在从多项式插值开始．假设要求(1)式中积分的值．可以选取$[a, b]$中的结点 x_0，x_1，…，x_n，建立如 6.1 节中那样的拉格朗日插值过程．定义

$$\ell_i(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j} \quad (0 \leqslant i \leqslant n)$$

这些是**基本插值多项式**．在结点上插值 f 的次数最多是 n 次的多项式是

$$p(x) = \sum_{i=0}^{n} f(x_i)\ell_i(x) \tag{2}$$

然后，如前面提到的那样，我们简单地写出下式

$$\int_a^b f(x)\,dx \approx \int_a^b p(x)\,dx = \sum_{i=0}^{n} f(x_i)\int_a^b \ell_i(x)\,dx$$

用这种方法，我们得到一个可以用于任何 f 的公式．公式如下

$$\int_a^b f(x)\,dx \approx \sum_{i=0}^{n} A_i f(x_i) \tag{3}$$

其中

$$A_i = \int_a^b \ell_i(x)\,dx$$

如果结点是等距的，那么形如(3)式的公式称为**牛顿-科茨公式**．

7.2.2 梯形法则

480

最简单的情况是 $n=1$ 并且结点取为 $x_0=a$，$x_1=b$．于是基本的插值多项式是

$$\ell_0(x) = \frac{b-x}{b-a}, \ell_1(x) = \frac{x-a}{b-a}$$

从而，

$$A_0 = \int_a^b \ell_0(x)\,dx = \frac{1}{2}(b-a) = \int_a^b \ell_1(x)\,dx = A_1$$

相应的求积公式是

$$\int_a^b f(x)\,dx \approx \frac{b-a}{2}[f(a) + f(b)]$$

该公式称为**梯形法则**．对所有 $f \in \Pi_1$（即次数最多是 1 次的全体多项式）公式精确成立．此外，它的误差项是

$$-\frac{1}{12}(b-a)^3 f''(\xi)$$

其中 $\xi\in(a, b)$. 通过对多项式逼近中的误差项 $f(x)-p_1(x)=f''(\xi_x)(x-a)(x-b)/2$ 积分，再利用积分中值定理，可以确定梯形法则的误差项. 在7.4节中，作为龙贝格积分法的组成部分，我们将再次看到梯形法则.

如果划分区间$[a, b]$为：

$$a=x_0<x_1<\cdots<x_n=b$$

那么在每个子区间上可应用梯形法则. 这时结点未必是等距的. 这样，我们得到**复合梯形法则**：

$$\int_a^b f(x)\mathrm{d}x=\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i} f(x)\mathrm{d}x$$

$$\approx\frac{1}{2}\sum_{i=1}^{n}(x_i-x_{i-1})[f(x_{i-1})+f(x_i)]$$

(把单个区间上的积分公式应用在区间划分以后的每个子区间上，这样便得到一个**复合法则**.)如果用1次插值样条函数(即折线)替换被积函数 f，也会出现复合梯形法则.

对等间距 $h=(b-a)/n$ 及结点 $x_i=a+ih$，**复合梯形法则**具有形式

$$\int_a^b f(x)\mathrm{d}x\approx\frac{h}{2}[f(a)+2\sum_{i=1}^{n-1}f(a+ih)+f(b)]$$

481 或者

$$\int_a^b f(x)\mathrm{d}x\approx h\sum_{i=0}^{n}{}'' f(a+ih) \tag{4}$$

其中求和符号上的两撇表示求和式中的第一项和最后一项都被减半. 复合梯形法则的误差项是

$$-\frac{1}{12}(b-a)h^2 f''(\xi)$$

其中 $\xi\in(a,b)$. 对每个子区间上的误差项求和并利用以下事实：在$[a,b]$内存在一点ξ使得$f''(\xi)=(1/n)\sum_{i=1}^{n}f''(\xi_i)$，其中 $\xi_i\in(x_{i-1},x_i)$ 以及 $1/n=(b-a)/h$，即平均值，这样便得到总误差项.

例1 在牛顿-科茨方法中，如果取 $n=2$，$[a, b]=[0, 1]$，则得到另一个公式：

$$\int_0^1 f(x)\mathrm{d}x\approx\frac{1}{6}f(0)+\frac{2}{3}f\left(\frac{1}{2}\right)+\frac{1}{6}f(1) \tag{5}$$

利用(3)式推导出这个结果.

解 对结点0，1/2，1的三个基本多项式是

$$\ell_0(x)=2\left(x-\frac{1}{2}\right)(x-1)\quad \ell_1(x)=-4x(x-1)\quad \ell_2(x)=2x\left(x-\frac{1}{2}\right)$$

则

$$A_0=\int_0^1\ell_0(x)\mathrm{d}x=\frac{1}{6}$$

等等. ■

7.2.3 待定系数法

从公式(3)的推导过程中，我们立刻看出对于所有次数$\leqslant n$的多项式，(3)式是精确成立的．另一方面，假设给定公式(3)，我们也只能知道对所有次数$\leqslant n$的多项式它是精确成立的．试问下面的等式成立吗？

$$A_i = \int_a^b \ell_i(x)\mathrm{d}x$$

答案是肯定的，这是因为公式(3)可以正确地积分任何ℓ_j．因此，

$$\int_a^b \ell_j(x)\mathrm{d}x = \sum_{i=0}^n A_i\ell_j(x_i) = A_j$$

482

当然，这里我们用到基本多项式的两条基本性质：ℓ_j是次数至多是n次的多项式以及$\ell_i(x_j)=\delta_{ij}$．

刚才的讨论使得能够用**待定系数法**得到与(3)类似的公式．为了说明这一点，我们用这种方法推导出(5)式．寻找一个公式

$$\int_0^1 f(x)\mathrm{d}x \approx A_0 f(0) + A_1 f\left(\frac{1}{2}\right) + A_2 f(1)$$

它对于所有次数$\leqslant 2$的多项式是精确成立的．依次把多项式$f(x)=1$，x及x^2作为试用函数，得到

$$1 = \int_0^1 \mathrm{d}x = A_0 + A_1 + A_2$$

$$\frac{1}{2} = \int_0^1 x\mathrm{d}x = \frac{1}{2}A_1 + A_2$$

$$\frac{1}{3} = \int_0^1 x^2\mathrm{d}x = \frac{1}{4}A_1 + A_2$$

这三个方程的联立方程组的解是：$A_0=1/6$，$A_1=2/3$及$A_2=1/6$．因为公式是线性的，所以对任何二次多项式$f(x)=c_0+c_1x+c_2x^2$，它将产生积分的精确值．

7.2.4 辛普森法则

对任意区间$[a,b]$的类似计算可得到熟悉的**辛普森法则**：

$$\int_a^b f(x)\mathrm{d}x \approx \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] \tag{6}$$

从它的推导过程可知，对于所有次数$\leqslant 2$的多项式辛普森法则是精确成立的．出乎意料地是，对于所有次数$\leqslant 3$的多项式它也精确成立．（见习题7.1.2.）

与辛普森法则联系在一起的误差项是：

$$-\frac{1}{90}[(b-a)/2]^5 f^{(4)}(\xi)$$

其中$\xi\in(a,b)$．如习题7.6.5中所示的那样，利用佩亚诺核定理，很容易导出这个误差项．根据下述讨论，可以看出误差是$\mathcal{O}(h^5)$．如果$h=(b-a)/2$，则该数值积分公式具有如下形式

$$\int_a^{a+2h} f(x)\mathrm{d}x \approx \frac{h}{3}[f(a) + 4f(a+h) + f(a+2h)]$$

利用1.1节中的泰勒定理，上式右端项可写为

483

$$2hf(a)+2h^2f'(a)+\frac{4}{3}h^3f''(a)+\frac{2}{3}h^4f'''(a)+\frac{100}{3\cdot 5!}h^5f^{(4)}(a)+\cdots$$

接着设

$$F(x)=\int_a^x f(t)\mathrm{d}t$$

根据微积分基本定理，$F'=f$，再利用泰勒定理，可以把辛普森法则的左端写为$F(a+2h)$或者

$$2hf(a)+2h^2f'(a)+\frac{4}{3}h^3f''(a)+23h^4f'''(a)+\frac{32}{5!}h^5f^{(4)}(a)+\cdots$$

结合这两个展开式，有

$$\int_a^{a+2h}f(x)\mathrm{d}x=\frac{h}{3}[f(a)+4f(a+h)+f(a+2h)]-\frac{1}{90}h^5f^{(4)}(a)-\cdots$$

我们经常使用偶数个子区间上的**复合辛普森法则**．设n是偶数，并且令

$$x_i=a+ih\qquad h=(b-a)/n\qquad (0\leqslant i\leqslant n)$$

则有

$$\begin{aligned}\int_a^b f(x)\mathrm{d}x&=\int_{x_0}^{x_2}f(x)\mathrm{d}x+\int_{x_2}^{x_4}f(x)\mathrm{d}x+\cdots+\int_{x_{n-2}}^{x_n}f(x)\mathrm{d}x\\&=\sum_{i=1}^{n/2}\int_{x_{2i-2}}^{x_{2i}}f(x)\mathrm{d}x\end{aligned}$$

把辛普森法则(6)应用于每个子区间，得到公式

$$\int_a^b f(x)\mathrm{d}x\approx\frac{h}{3}\sum_{i=1}^{n/2}[f(x_{2i-2})+4f(x_{2i-1})+f(x_{2i})]$$

为了避免项的重复，这个公式的右端项可以计算如下：

$$\frac{h}{3}\left[f(x_0)+2\sum_{i=2}^{n/2}f(x_{2i-2})+4\sum_{i=1}^{n/2}f(x_{2i-1})+f(x_n)\right]$$

公式的误差项是：

$$-\frac{1}{180}(b-a)h^4f^{(4)}(\xi)$$

其中$\xi\in(a,b)$．

7.2.5 一般积分公式

推导牛顿-科茨公式的过程可以用来产生下列更一般的积分公式

484

$$\int_a^b f(x)w(x)\mathrm{d}x\approx\sum_{i=0}^{n}A_if(x_i)$$

其中w可以是任何固定的**权函数**．必要的改进仅仅是把

$$A_i=\int_a^b \ell_i(x)w(x)\mathrm{d}x$$

代入(3)式中．下面的例题将说明这一点．

例2 求出一个公式

$$\int_{-\pi}^{\pi} f(x)\cos x \mathrm{d}x \approx A_0 f\left(-\frac{3}{4}\pi\right) + A_1 f\left(-\frac{1}{4}\pi\right) + A_2 f\left(\frac{1}{4}\pi\right) + A_3 f\left(\frac{3}{4}\pi\right)$$

当 f 是一个三次多项式时它是精确成立的.

解 显然，一个三次多项式是 4 个单项式 1，x，x^2 及 x^3 的线性组合. 因而，通过代入 $f(x)=x^j$，$(0\leqslant j\leqslant 3)$以及求解得到的 4 个线性方程，可以确定其系数. 根据对称性，$A_0=A_3$，$A_1=A_2$，这样，结果简化为

$$0=\int_{-\pi}^{\pi} 1\cos x \mathrm{d}x = 2A_0 + 2A_1$$

$$-4\pi=\int_{-\pi}^{\pi} x^2\cos x \mathrm{d}x = 2A_0\left(\frac{3}{4}\pi\right)^2 + 2A_1\left(\frac{1}{4}\pi\right)^2$$

解是 $A_1=A_2=-A_0=-A_3=4/\pi$，则公式为

$$\int_{-\pi}^{\pi} f(x)\cos x \mathrm{d}x \approx \frac{4}{\pi}\left[-f\left(-\frac{3}{4}\pi\right) + f\left(-\frac{1}{4}\pi\right) + f\left(\frac{1}{4}\pi\right) - f\left(\frac{3}{4}\pi\right)\right]$$ ■

7.2.6 区间变换

经过变量的线性变换我们可以从某一个区间上的数值积分公式导出任一其他区间上的公式. 如果对于某些次数的多项式第一个公式是精确成立的，那么第二个公式对于它们也是精确成立的. 现在考虑一下如何做到这一点.

假设给定数值积分公式：

$$\int_c^d f(t)\mathrm{d}t \approx \sum_{i=0}^{n} A_i f(t_i) \tag{7}$$

我们并不介意公式来自哪里；然而，要假设对于所有次数$\leqslant m$ 的多项式它是精确成立的. 如果需要另外某个区间例如$[a, b]$上的公式，我们首先定义 t 的线性函数λ，使得当 t 在$[c, d]$中变动时函数 $\lambda(t)$在$[a, b]$中变动. 函数 λ 由下式显式给出

$$\lambda(t) = \frac{b-a}{d-c}t + \frac{ad-bc}{d-c} \tag{8}$$

485

接着，对积分

$$\int_a^b f(x)\mathrm{d}x$$

作变量替换 $x=\lambda(t)$. 于是 $\mathrm{d}x=\lambda'(t)\mathrm{d}t=(b-a)(d-c)^{-1}\mathrm{d}t$，并且有

$$\int_a^b f(x)\mathrm{d}x = \frac{b-a}{d-c}\int_c^d f(\lambda(t))\mathrm{d}t$$

$$\approx \frac{b-a}{d-c}\sum_{i=0}^{n} A_i f(\lambda(t_i))$$

因此，有

$$\int_a^b f(x)\mathrm{d}x \approx \frac{b-a}{d-c}\sum_{i=0}^{n} A_i f\left(\frac{b-a}{d-c}t_i + \frac{ad-bc}{d-c}\right) \tag{9}$$

可以看出，由于λ是线性的，如果 f 是一个多项式，则 $f(\lambda(t))$是 t 的多项式，并且它们的次数相同. 因此，对于 m 次的多项式，新的公式也是精确成立的. 在辛普森法则中，该过程利

用 $\lambda(t)=(b-a)t+a$ 可由公式(5)得到公式(6).

7.2.7 误差分析

为了估计数值积分公式(3)中的误差，我们要用到多项式插值中的误差项. 回顾 6.1 节中(13)式，如果 p 是在点 x_0，x_1，…，x_n 上插值 f 的次数$\leqslant n$ 的多项式，而且 $f^{(n+1)}$连续，则有

$$f(x)-p(x)=\frac{1}{(n+1)!}f^{(n+1)}(\xi_x)\prod_{i=0}^{n}(x-x_i) \tag{10}$$

因而，

$$\int_a^b f(x)\mathrm{d}x-\sum_{i=0}^{n}A_i f(x_i)=\frac{1}{(n+1)!}\int_a^b f^{(n+1)}(\xi_x)\prod_{i=0}^{n}(x-x_i)\mathrm{d}x \tag{11}$$

如果在$[a, b]$上 $|f^{(n+1)}(x)|\leqslant M$，则有

$$\left|\int_a^b f(x)\mathrm{d}x-\sum_{i=0}^{n}A_i f(x_i)\right|\leqslant\frac{M}{(n+1)!}\int_a^b\prod_{i=0}^{n}|x-x_i|\mathrm{d}x \tag{12}$$

使不等式右端项尽可能小的结点可选取为

486

$$x_i=\frac{a+b}{2}+\frac{b-a}{2}\cos\left[\frac{(i+1)\pi}{n+2}\right]\quad(0\leqslant i\leqslant n)$$

如果区间是$[-1, 1]$，则这些结点具有更简单的形式

$$x_i=\cos\left[\frac{(i+1)\pi}{n+2}\right]\quad(0\leqslant i\leqslant n)$$

它们是下列函数的零点

$$U_{n+1}(x)=\frac{\sin[(n+2)\theta]}{\sin\theta}\quad(x=\cos\theta) \tag{13}$$

函数 U_{n+1}称为**第二类切比雪夫多项式**；如(13)式中所定义的那样，$U_{n+1}(x)$不是首一多项式；即它的首项系数不是 1. 事实上，U_{n+1}中 x^{n+1}的系数是 2^{n+1}. 因此，

$$2^{-(n+1)}U_{n+1}=(x-x_0)(x-x_1)\cdots(x-x_n) \tag{14}$$

其中 $x_i=\cos[(i+1)\pi/(n+2)]$. 然后，经计算得

$$\int_{-1}^{1}|(x-x_0)(x-x_1)\cdots(x-x_n)|\mathrm{d}x=2^{-n} \tag{15}$$

因而，利用这些结点以及适当的系数 A_i，我们有

$$\left|\int_{-1}^{1}f(x)\mathrm{d}x-\sum_{i=0}^{n}A_i f(x_i)\right|\leqslant\frac{M}{(n+1)!2^n} \tag{16}$$

定理 1(极值性质定理，第二类切比雪夫多项式) *在 n 次首一多项式 p 中，使得 $\int_{-1}^{1}|p(x)|\mathrm{d}x$ 最小的多项式是 $2^{-n}U_n$.*

证明 首先，证明正交关系

$$\int_{-1}^{1}U_m(x)\mathrm{sign}[U_n(x)]\mathrm{d}x=0\quad(0\leqslant m\leqslant n)$$

设 I 表示这个积分. 作变量替换 $\cos\theta=x$ 得

$$I=\int_0^\pi\frac{\sin(m+1)\theta}{\sin\theta}\mathrm{sign}\left[\frac{\sin(n+1)\theta}{\sin\theta}\right]\sin\theta\mathrm{d}\theta$$

$$= \sum_{k=0}^{n}(-1)^k\int_{k\varphi}^{(k+1)\varphi}\sin(m+1)\theta\mathrm{d}\theta \quad \varphi=\frac{\pi}{(n+1)}$$

$$= (m+1)^{-1}\sum_{k=0}^{n}(-1)^{k+1}[\cos(m+1)(k+1)\varphi-\cos(m+1)k\varphi]$$

487

令 $\alpha=(m+1)\varphi+\pi$. 然后用 $\mathrm{Re}(x)$表示 x 的实部，我们有

$$(m+1)I = \sum_{k=0}^{n}[\cos(k+1)\alpha+\cos k\alpha]$$

$$= \mathrm{Re}\left\{\sum_{k=0}^{n}[\mathrm{e}^{\mathrm{i}\alpha(k+1)}+\mathrm{e}^{\mathrm{i}\alpha k}]\right\}$$

$$= \mathrm{Re}\left\{\frac{\mathrm{e}^{\mathrm{i}\alpha(n+2)}-\mathrm{e}^{\mathrm{i}\alpha}+\mathrm{e}^{\mathrm{i}\alpha n}-1}{\mathrm{e}^{\mathrm{i}\alpha}-1}\right\}$$

$$= \frac{\mathrm{Re}[(\mathrm{e}^{-\mathrm{i}\alpha}-1)(\mathrm{e}^{\mathrm{i}\alpha(n+2)}-\mathrm{e}^{\mathrm{i}\alpha}+\mathrm{e}^{\mathrm{i}\alpha n}-1)]}{|\mathrm{e}^{\mathrm{i}\alpha}-1|^2}$$

这里分子是

$$\mathrm{Re}(\mathrm{e}^{\mathrm{i}\alpha n}-\mathrm{e}^{\mathrm{i}\alpha(n+2)}) = \cos n\alpha-\cos(n+2)\alpha$$

$$= \cos[(n+1)\alpha-\alpha]-\cos[(n+1)\alpha+\alpha]$$

$$= \cos(k\pi-\alpha)-\cos(k\pi+\alpha) = 0$$

其中 $k=m+n+2$. 为了完成证明，令 p 是n 次首一多项式. 它也可以表示为

$$p = 2^{-n}U_n+a_{n-1}U_{n-1}+\cdots+a_0U_0$$

从而，根据正交关系，有

$$\int_{-1}^{1}|p|\mathrm{d}x \geqslant \int_{-1}^{1}p\mathrm{sign}[U_n]\mathrm{d}x = 2^{-n}\int_{-1}^{1}U_n\mathrm{sign}[U_n]\mathrm{d}x$$

$$= 2^{-n}\int_{-1}^{1}|U_n|\mathrm{d}x$$

■

有关数值积分的主题可进一步参见 Davis and Rabinowitz[1984]、Krylov[1962]以及 Ghizetti and Ossiccini[1970].

习题 7.2

1. 推导出基于结点 0，1/3，2/3，1 的 $\int_0^1 f(x)\mathrm{d}x$ 的牛顿-科茨公式.
2. 证明(不要利用它的误差项)：辛普森法则(6)式正确地积分所有三次多项式.
3. 用一个适当的变量替换从(5)式得到(6)式.
4. 证明：下列公式对于次数≤4 的多项式是精确成立的：

$$\int_0^1 f(x)\mathrm{d}x \approx \frac{1}{90}\left[7f(0)+32f\left(\frac{1}{4}\right)+12f\left(\frac{1}{2}\right)+32f\left(\frac{3}{4}\right)+7f(1)\right]$$

5. (续)根据上题中的公式，对 $\int_a^b f(x)\mathrm{d}x$ 推导一个公式使得它对于所有四次多项式是精确成立的.

488

6. (续)把上题中的公式应用于

$$\int_0^1\frac{\mathrm{d}t}{t+1}$$

近似计算 ln 2. 把你的答案与正确值作比较并且计算误差.

7. 利用课本中的级数，计算$\int_0^1 \mathrm{e}^{x^2}\mathrm{d}x$，要求达到 8 位小数的精确度.

8. 求出公式

$$\int_0^1 f(x)\mathrm{d}x \approx A_0 f(0) + A_1 f(1)$$

使得它对所有形为 $f(x)=ae^x+b\cos(\pi x/2)$ 的函数是精确成立的.

9. 求出下列形式的公式：

$$\int_0^{2\pi} f(x)\mathrm{d}x = A_1 f(0) + A_2 f(\pi)$$

使得它对任何具有形式

$$f(x) = a + b\cos x$$

的函数是精确成立的. 证明所得到的公式对任何形为

$$f(x) = \sum_{k=0}^{n}[a_k\cos(2k+1)x + b_k\sin kx]$$

的函数也是精确成立的.

10. 利用拉格朗日插值多项式推导出公式

$$\int_0^1 f(x)\mathrm{d}x \approx Af\left(\frac{1}{3}\right) + Bf\left(\frac{2}{3}\right)$$

把该公式变换为区间$[a, b]$上的积分公式.

11. 利用在 x_1 和 x_2 上插值 $f(x)$ 的最低阶多项式，推导出数值积分公式

$$\int_{x_0}^{x_3} f(x)\mathrm{d}x$$

不要假设等距点. 这里 $x_0<x_1<x_2<x_3$.

12. 按照 $f(0)$，$f(2)$ 及 $f(4)$，推导出近似下列积分的公式：

$$\int_1^3 f(x)\mathrm{d}x$$

它对于 Π_2 中所有 f 应该是精确成立的.

13. 确定 A，B 及 C 的值使得公式

$$\int_0^2 xf(x)\mathrm{d}x \approx Af(0) + Bf(1) + Cf(2)$$

对于所有次数尽可能高的多项式是精确成立的. 试问最大次数是多少？

14. 基于结点 -2，-1 及 0 上拉格朗日插值多项式推导

$$\int_0^1 f(x)\mathrm{d}x$$

489

的牛顿-科茨公式. 当 $f(x)=\sin \pi x$ 时，利用这个结果计算积分的值.

15. 利用级数，计算

$$\int_0^{10^{-2}} \left(\frac{\sin x}{x}\right)\mathrm{d}x$$

要求达到 7 位小数精度.

16. 我们打算用$\int_0^1 p(x)\mathrm{d}x$ 作为$\int_0^1 f(x)\mathrm{d}x$ 的估计值，其中 p 是$[0,1]$中结点 $x_0, x_1, \cdots, x_n$ 上插值 f 的 n 次多项式. 假设在$[0,1]$上 $|f^{(n+1)}(x)|<M$. 如果不知道结点的位置，对误差 $\left|\int_0^1 f(x)\mathrm{d}x - \int_0^1 p(x)\mathrm{d}x\right|$ 能给出什么样的上界？试问你能找到最佳的上界吗？

17. 基于简单的右边矩形法则：

$$\int_0^1 f(x)\mathrm{d}x \approx f(1)$$

确定复合数值积分法则. 假设 $a=x_0<x_1<\cdots<x_n=b$ 不等距.

18. 基于中点法则

$$\int_{-1}^1 f(x)\mathrm{d}x \approx 2f(0)$$

推导出 $\int_a^b f(x)\mathrm{d}x$ 的复合法则. 分别给出等距结点和不等距结点的公式.

19. (续)区间$[x_{i-1}, x_{i+1}]$上的**中点法则**如下:

$$\int_{x_{i-1}}^{x_{i+1}} f(x)\mathrm{d}x = (x_{i+1}-x_{i-1})f(x_i)$$

利用等间距 $h=(b-a)/n$, $x_i=a+ih$, $i=0, 1, 2, \cdots, n$(n 是偶数), 确定区间$[a, b]$上的复合中点法则.

20. 基于高斯求积法则

$$\int_{-1}^1 f(x)\mathrm{d}x \approx f\left(-\frac{1}{\sqrt{3}}\right)+f\left(\frac{1}{\sqrt{3}}\right)$$

确定 $\int_a^b f(x)\mathrm{d}x$ 的积分法则.

21. 对于 $n=2$ 和$[a, b]=[0, 1]$, 有两个牛顿-科茨公式; 即

$$\int_0^1 f(x)\mathrm{d}x \approx af(0)+bf\left(\frac{1}{2}\right)+cf(1)$$

$$\int_0^1 f(x)\mathrm{d}x \approx \alpha f\left(\frac{1}{4}\right)+\beta f\left(\frac{1}{2}\right)+\gamma f\left(\frac{3}{4}\right)$$

试问哪一个更好?

22. 试问是否存在一个形为

$$\int_0^1 f(x)\mathrm{d}x \approx \alpha[f(x_0)+f(x_1)]$$

的公式可正确地积分所有二次多项式?

23. 证明: 如果公式

$$\int_{-1}^1 f(x)\mathrm{d}x \approx \sum_{i=0}^n A_i f(x_i) \quad (n\text{ 是偶数})$$

490

对所有 n 次多项式是精确成立的, 并且如果结点是关于原点对称地放置的, 那么公式对于所有 $n+1$ 次多项式是精确成立的.

24. 设 n 是偶数. 说明怎样用最小的附加工作量, 从 n 个等距结点的情况计算出 $2n$ 个等距结点的复合辛普森法则.

25. 在例 2 中, 对称性被用于简化计算. 给出 $A_0=A_3$ 和 $A_1=A_2$ 的一种证明.

26. 证明: 由下式递归生成**第二类切比雪夫多项式**:

$$\begin{cases} U_0(x)=1 \quad U_1(x)=2x \\ U_{n+1}=2xU_n-U_{n-1} \quad (n\geqslant 1) \end{cases}$$

27. (续)证明下列正交关系:

$$\int_{-1}^1 U_n(x)U_m(x)\sqrt{1-x^2}\,\mathrm{d}x=\delta_{nm}\frac{\pi}{2}$$

28. (续)证明第一类和第二类切比雪夫多项式之间的关系: $T'_n=nU_{n-1}$.

29. 建立梯形法则和复合梯形法则的误差项.

30. 给出建立辛普森法则和复合辛普森法则误差项过程的细节.
31. 利用梯形法则近似计算

$$\int_1^2 (x+e^{-x^2})\mathrm{d}x$$

为了至少达到$(1/2)\times 10^{-7}$的精度，确定所需要的子区间的最少个数.

计算机习题 7.2

1. 编写一个计算机程序，用适当的泰勒级数求和计算 $\int_0^x e^{-t^2}\mathrm{d}t$，要求直到级数个别项的量值小于 10^{-8} 为止. 对于 $x=0.0, 0.1, 0.2, \cdots, 1.0$，通过计算该积分的值检验你的程序.
2. 编写一个用 $\int_a^b S(x)\mathrm{d}x$ 估计 $\int_a^b f(x)\mathrm{d}x$ 的计算机程序，其中 S 是结点 $a+ih$ 上插值 f 的自然三次样条，这里 $0\leqslant i\leqslant n$，$h=(b-a)/n$. 首先从 6.4 节中(7)式开始，得到积分

$$\int_{t_0}^{t_n} S(x)\mathrm{d}x$$

的一个公式. 然后写出计算它的子程序. 用下列众所周知的积分测试你的程序：

a. $\dfrac{4}{\pi}\int_0^1 (1+x^2)^{-1}\mathrm{d}x$

b. $\dfrac{1}{\ln 3}\int_1^3 x^{-1}\mathrm{d}x$

在这两个例子中分别取 $n=4, 8, 16$.（见计算机习题 6.4.3.）

3. 用一个符号操作程序，完成下列练习：

[491] a. 求出不定积分 $\int (\cos x)^{-14}\mathrm{d}x$.

b. 求出定积分 $\int_0^1 \log(\log x)\mathrm{d}x$.

c. 求 $\int_0^1 \sqrt{1+\sin^3 x}\,\mathrm{d}x$ 的数值.

7.3 高斯求积

在上一节中，我们看到如何产生下列形式的求积公式

$$\int_a^b f(x)\mathrm{d}x \approx \sum_{i=0}^{n} A_i f(x_i) \tag{1}$$

这个公式对于次数$\leqslant n$ 的多项式是**精确成立的**. 在这些公式中，结点 $x_0, x_1, x_2, \cdots, x_n$ 是事先选定的. 一旦结点选定，根据对于所有 $f\in \Pi_n$ 公式(1)必须是等式的需要，就可以唯一地确定其系数.

自然要问在公式(1)中结点的某种选法是否比另一些选法更好. 例如，可能存在一组特殊的结点使得系数 A_i 彼此相等. 如果 $A_i=c$，$0\leqslant i\leqslant n$，使用(1)式就会减少算术运算，因为这时公式具有更简单的形式

$$\int_a^b f(x)\mathrm{d}x \approx c\sum_{i=0}^{n} f(x_i) \tag{2}$$

（它使乘法运算的次数从 $n+1$ 次减少到 1 次.）只有 $n=0, 1, 2, 3, 4, 5, 6, 8$ 时才存在(2)式类型的公式. 它们称为**切比雪夫求积式**. $n=4$ 对应的公式是

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx \frac{2}{5}[f(-\alpha)+f(-\beta)+f(0)+f(\beta)+f(\alpha)] \tag{3}$$

其中

$$\alpha=\sqrt{(5+\sqrt{11})/12}\approx 0.832\ 497\ 487\ 000\ 982$$

$$\beta=\sqrt{(5-\sqrt{11})/12}\approx 0.374\ 541\ 409\ 553\ 581$$

可用待定系数法来得到结点 α 和 β. 而且可以说明对于次数 $\leqslant 5$ 的多项式该公式是精确成立的，还可以给出具有相同系数的稍显复杂的公式. 举例如下：

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx \frac{\pi}{n}\sum_{i=1}^{n} F\left(\cos\frac{2i-1}{2n}\pi\right) \tag{4}$$

其中 $F(x)=f(x)\sqrt{1-x^2}$. 该公式称为**埃尔米特求积式**. 对于一个特定的 $2n$ 维线性空间它是精确成立的，即

$$G=\{pw: p\in \Pi_{2n-1}\} \qquad w(x)=(1-x^2)^{-1/2}$$

492

使用该公式时将不可能减少计算.（为什么?）另一方面，我们希望得到一个公式它仅有 n 次 p 的赋值而在一个 $2n$ 维线性空间上是精确成立的. 为了追求这个目标，我们系统地引出高斯求积公式.

7.3.1 高斯求积公式

可以对稍微更一般形式的求积法则来阐明这个理论，一般形式是

$$\int_a^b f(x)w(x)\mathrm{d}x \approx \sum_{i=0}^{n} A_i f(x_i) \tag{5}$$

其中 w 是给定的正的**权函数**. 自然地，$w(x)\equiv 1$ 是一种特别重要的情况. 我们假设对于 $f\in \Pi_n$ 公式(5)是精确成立的，从上一节内容可知，假设成立当且仅当

$$A_i=\int_a^b w(x)\prod_{\substack{j=0\\ j\neq i}}^{n}\frac{x-x_j}{x_i-x_j}\mathrm{d}x \tag{6}$$

因为有 $n+1$ 个系数 A_i 和 $n+1$ 个结点 x_i，这里的处理方法不预先限制结点，所以我们猜测会找到形如(5)式并且对于次数 $\leqslant 2n+1$ 的多项式是精确成立的求积公式. 现在我们将指出结果确实如此.

Carl Friedrich Gauss(1777—1855)给出了这样的想法，利用结点的可变性迫使求积公式(5)和(6)对于所有 $2n+1$ 次多项式是精确成立的. 下面的定理揭示了结点所处的位置.

定理 1(高斯求积定理) *设 w 是正的权函数，q 是一个 $n+1$ 次非零多项式并且与 Π_n 是 w* ***正交的**，也就是对任意 $p\in\Pi_n$ 都有*

$$\int_a^b q(x)p(x)w(x)\mathrm{d}x=0 \tag{7}$$

若 x_0，x_1，…，x_n 是 q 的零点，则具有(6)式中给定系数的求积公式(5)对于所有 $f\in\Pi_{2n+1}$ 是精确成立的.

证明 设 $f\in\Pi_{2n+1}$. 用 q 除 f，得到商式 p 和余式 r，则有

$$f=qp+r \qquad (p,r\in\Pi_n)$$

因而，$f(x_i)=r(x_i)$. 利用(7)式，再根据(5)式对 Π_n 中的元素是精确成立的，我们有

$$\int_a^b fw\mathrm{d}x=\int_a^b rw\mathrm{d}x=\sum_{i=0}^{n}A_ir(x_i)=\sum_{i=0}^{n}A_if(x_i)$$ ■

493

由此可得 q 的根是单根并且都落在区间$[a, b]$的内部.（特别地，它们都是实数.）这一点直接由下面的定理得到.

定理 2(符号变化次数定理) 设 w 是 $C[a, b]$中正的权函数，并且 f 是 $C[a, b]$中与 Π_n 是 w 正交的非零元，则 f 在(a, b)上至少变号 $n+1$ 次.

证明 因为 $1\in\Pi_n$，所以有 $\int_a^b f(x)w(x)\mathrm{d}x=0$，这表明 f 至少变号一次. 假设 f 只变号 r 次，$r\leqslant n$. 选取点 t_i 使得

$$a=t_0<t_1<t_2<\cdots<t_r<t_{r+1}=b$$

而且在下列每个区间上 f 不变号.

$$(t_0,t_1),(t_1,t_2),\cdots,(t_r,t_{r+1})$$

多项式

$$p(x)=\prod_{i=1}^{r}(x-t_i)$$

具有同号性质，从而 $\int_a^b f(x)p(x)w(x)\mathrm{d}x\neq 0$. 因为 $p\in\Pi_n$，所以与原假设矛盾. ■

如果权函数是 $w(x)=1$ 并且区间是$[-1, 1]$，我们得到高斯原创的研究工作. 这种情况下对应于 $n=1$ 和 $n=4$ 的两个公式如下：

$$\int_{-1}^{1} f(x)\mathrm{d}x\approx f(-\alpha)+f(\alpha) \tag{8}$$

其中，$\alpha=1/\sqrt{3}$，以及

$$\int_{-1}^{1} f(x)\mathrm{d}x\approx A_0f(x_0)+A_1f(x_1)+\cdots+A_4f(x_4) \tag{9}$$

其中

$$-x_0=x_4=\frac{1}{3}\sqrt{5+2\sqrt{10/7}}\approx 0.906\,179\,845\,938\,664$$

$$-x_1=x_3=\frac{1}{3}\sqrt{5-2\sqrt{10/7}}\approx 0.538\,469\,310\,105\,683$$

$$x_2=0.0$$

$$A_0=A_4=0.3(0.7+5\sqrt{0.7})/(2+5\sqrt{0.7})\approx 0.236\,926\,885\,056\,189$$

$$A_1=A_3=0.3(-0.7+5\sqrt{0.7})/(-2+5\sqrt{0.7})\approx 0.478\,628\,670\,499\,366$$

494

$$A_2=128/225\approx 0.568\,888\,888\,888\,889$$

通过使用确切的结点和系数，可以计算这些积分使其达到任意要求的精度. 事实上，结点 x_i 分别是勒让德多项式 $p_2(x)=\frac{1}{2}(3x^2-1)$和 $p_5(x)=\frac{1}{8}(63x^5-70x^3+15x)$的根，它们在 6.8 节的例 2 中给出.

各种类型的很多求积公式的结点 x_i 和系数 A_i 可以在 Abramowitz and Stegun[1964]这样的手册中查到. 手册中列出 n 的值是 $n=1$, 2, …, 9, 11, 15, 19, 23, 31, 39, 47, 63, 79, 95. 对于好的被积函数，使用高斯公式只需计算几个函数的值便可得到合理的精度，使用一些高阶公式还可得到更高的精度.

结合前面的 5 点高斯求积公式，给出一个简短的伪代码用于近似计算积分

$$\int_a^b f(x)\mathrm{d}x$$

这里我们把区间变为 7.2 节(9)式中所述的那样，并且利用(8)式中的对称性.

```
input
    x0 ← 0
    x1 ← 0.538 469 310 105 683
    x2 ← 0.906 179 845 938 664
    A0 ← 0.568 888 888 888 889
    A1 ← 0.478 628 670 499 366
    A2 ← 0.236 926 885 056 189
u ← (a+b)/2
S ← A0 f(u)
for i = 1 to 2 do
    u ← ((b-a)xi + a + b)/2
    v ← ((a-b)xi + a + b)/2
    S ← S + Ai[f(u) + f(v)]
end do
S ← (b-a)S/2
output S
```

一旦结点 x_i 被确定，高斯公式中系数 A_i 的计算就按照非高斯公式中所示的那样进行. 结点依次是某一多项式的根，这个多项式与 $n+1$ 有关，我们把它记为 q_{n+1}. 下列两个条件唯一地定义多项式 q_{n+1}：

1. q_{n+1}是一个 $n+1$ 次首一多项式.

2. q_{n+1}与 Π_n是 w 正交的.

术语**首一**指的是 q_{n+1}中 x^{n+1}的系数是 1. w **正交性**的条件是

$$\int_a^b q_{n+1}(x)p(x)w(x)\mathrm{d}x = 0,\text{对所有 } p \in \Pi_n$$

495

这些所谓的**正交多项式**在数学的很多分支中都非常有用，并且它们有许多熟知的性质. 如 6.8 节讨论的那样，它们可由一个递推算法生成.

例 1 当$[a, b]=[-1, 1]$, $w(x)=1$, $n=2$ 时，求出高斯求积法则.

解 根据 6.8 节定理 5，并且如该节例 2 中给出的那样，可以用递推方法确定正交多项式. 因此，

$$q_0(x) = 1$$
$$q_1(x) = x$$
$$q_2(x) = x^2 - \frac{1}{3}$$
$$q_3(x) = x^3 - \frac{3}{5}x$$

q_3 的根 0 和$\pm\sqrt{3/5}$是下列求积公式中的结点

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx \frac{5}{9}f\left(-\sqrt{\frac{3}{5}}\right)+\frac{8}{9}f(0)+\frac{5}{9}f\left(\sqrt{\frac{3}{5}}\right)$$

通过待定系数法(见习题 7.3.21)，可以求出常数 5/9 和 8/9. ■

7.3.2 收敛性和误差分析

接下来讨论高斯求积公式的一些令人称道的特性.

引理 1(高斯求积公式引理) 在高斯求积公式中，它的系数是正的，而且它们的和是$\int_a^b w(x)\mathrm{d}x$.

证明 固定 n，如定理 1 中那样，令 q 是 $n+1$ 次多项式并且与 Π_n 是 w 正交的. q 的零点记为 x_0，x_1，…，x_n，它们是高斯公式(5)中的结点. 对于某一固定 j，设 p 是多项式 $q(x)/(x-x_j)$. 因为 p^2 的次数最多是 $2n$，所以高斯公式对于它是精确成立的. 因而，

$$0<\int_a^b p^2(x)w(x)\mathrm{d}x=\sum_{i=0}^{n}A_i p^2(x_i)=A_j p^2(x_j)$$

由此可知，$A_j>0$. 因为高斯公式对于 $f(x)\equiv 1$ 是精确成立的，所以

496

$$\int_a^b w(x)\mathrm{d}x=\sum_{i=0}^{n}A_i$$ ■

斯蒂尔切斯的一个卓越的定理建立了当 $n\to\infty$时高斯公式的收敛性. 给定 w 和$[a, b]$，对每个 $n\in\{0, 1, 2, \cdots\}$我们都有一个高斯求积公式. 它们由下列定理给出.

定理 3(高斯求积收敛性定理) 若 f 在$[a, b]$上连续，则当 $n\to\infty$时近似公式

$$\int_a^b f(x)w(x)\mathrm{d}x \approx \sum_{i=0}^{n}A_{ni}f(x_{ni}) \quad (n\geqslant 0) \tag{10}$$

收敛于积分.

证明 设$\epsilon>0$. 根据魏尔斯特拉斯逼近定理(6.1 节定理 8)，存在多项式 p 使其在$[a, b]$上$|f(x)-p(x)|<\epsilon$. 对于任一整数 n，如果 $2n$ 大于 p 的次数，那么 n 次高斯公式能正确积分 p. 利用基本不等式高斯求积公式引理 1，我们有

$$\begin{aligned}
&\left|\int_a^b f(x)w(x)\mathrm{d}x-\sum_{i=0}^{n}A_{ni}f(x_{ni})\right| \\
\leqslant &\left|\int_a^b f(x)w(x)\mathrm{d}x-\int_a^b p(x)w(x)\mathrm{d}x\right|+\left|\sum_{i=0}^{n}A_{ni}p(x_{ni})-\sum_{i=0}^{n}A_{ni}f(x_{ni})\right| \\
\leqslant &\int_a^b |f(x)-p(x)|\,w(x)\mathrm{d}x+\sum_{i=0}^{n}A_{ni}\,|p(x_{ni})-f(x_{ni})| \\
\leqslant &\,\epsilon\int_a^b w(x)\mathrm{d}x+\epsilon\sum_{i=0}^{n}A_{ni}=2\epsilon\int_a^b w(x)\mathrm{d}x
\end{aligned}$$ ■

定理 4(带误差项的高斯公式定理) 考虑带有误差项的高斯公式

$$\int_a^b f(x)w(x)\mathrm{d}x=\sum_{i=0}^{n-1}A_i f(x_i)+E$$

对 $f\in C^{2n}[a,b]$，我们有

$$E=\frac{f^{(2n)}(\xi)}{(2n)!}\int_a^b q^2(x)w(x)\mathrm{d}x$$

其中 $a<\xi<b$，$q(x)=\prod_{i=0}^{n-1}(x-x_i)$.

证明 根据埃尔米特插值(6.3 节)，存在一个次数最多是 $2n-1$ 次的多项式 p，使得

$$p(x_i)=f(x_i)\quad p'(x_i)=f'(x_i)\quad (0\leqslant i\leqslant n-1)$$

497

正如 6.3 节定理 2 中给出的那样，这个插值的误差公式是

$$f(x)-p(x)=f^{(2n)}(\zeta(x))q^2(x)/(2n)! \tag{11}$$

由此可得

$$\int_a^b f(x)w(x)\mathrm{d}x-\int_a^b p(x)w(x)\mathrm{d}x=\frac{1}{(2n)!}\int_a^b f^{(2n)}(\zeta(x))q^2(x)w(x)\mathrm{d}x$$

利用 p 的次数最多是 $2n-1$，可以看出

$$\int_a^b p(x)w(x)\mathrm{d}x=\sum_{i=0}^{n-1}A_ip(x_i)=\sum_{i=0}^{n-1}A_if(x_i)$$

此外，可以利用积分中值定理给出

$$\int_a^b f^{(2n)}(\zeta(x))q^2(x)w(x)\mathrm{d}x=f^{(2n)}(\xi)\int_a^b q^2(x)w(x)\mathrm{d}x$$

这里需要用到 $f^{(2n)}(\zeta(x))$的连续性，从(11)式可知这一点成立. 简单地代入后便可得到所要求的误差公式. ■

有关高斯求积公式进一步的参考文献还有 Krylov[1962]、Davis and Rabinowitz[1956，1984]、Stroud and Secrest[1966]以及 Abramowitz and Stegun[1956，1964].

习题 7.3

1. 在(2)式中，令 $n=1$，$a=0$，$b=1$. 找出对于 $f\in\Pi_3$ 精确成立的公式的所有情况.
2. (续)对 $n=2$ 的情况，求解上题.
3. 推导下列高斯求积法则：

$$\int_{-1}^1 f(x)\mathrm{d}x\approx A_0f(x_0)+A_1f(x_1)+A_2f(x_2)+A_3f(x_3)$$

其中

$$A_0=A_2=\frac{1}{2}\left(1+\frac{1}{6}\sqrt{10/3}\right)\approx 0.347\ 854\ 845\ 137\ 454$$

$$A_1=A_3=\frac{1}{2}\left(1-\frac{1}{6}\sqrt{10/3}\right)\approx 0.652\ 145\ 154\ 862\ 546$$

$$-x_0=x_2=\sqrt{\frac{1}{7}(3-4\sqrt{0.3})}\approx 0.861\ 136\ 311\ 594\ 052$$

$$-x_1=x_3=\sqrt{\frac{1}{7}(3+4\sqrt{0.3})}\approx 0.339\ 981\ 043\ 584\ 845\ 6$$

4. 证明(8)式作为高斯公式的正确性.
5. 证明(9)式中的结点是正确的.

498

6. 通过表达式

$$q(x) = x^{n+1} + c_1 x^n + \cdots + c_{n+1}$$

以及附加条件 $\int_a^b q(x)x^k w(x)\mathrm{d}x = 0$，$0 \leqslant k \leqslant n$，我们可以确定一个与 Π_n 正交的 $n+1$ 次多项式 q. 则它产生的含有 $n+1$ 个未知量 c_1，c_2，…，c_{n+1} 的 $n+1$ 个方程的方程组是可解的. 执行上述过程可以求出习题 7.3.3 中所需要的 q_5. 你认为这是一个获得 q 的最好方法吗?

7. a. 求出下列形式的一个公式

$$\int_0^1 xf(x)\mathrm{d}x \approx \sum_{i=0}^n A_i f(x_i)$$

其中 $n=1$，这个公式对所有三次多项式都精确成立.

b. 对 $n=2$，重复上题使得公式在 Π_5 上精确成立.

8. a. 确定 A_i 和 x_i 的适当值使得求积公式

$$\int_{-1}^1 x^2 f(x)\mathrm{d}x \approx \sum_{i=0}^n A_i f(x_i)$$

对于任意三次多项式 f 是精确成立的，取 $n=1$.

b. 当 f 是任意五次多项式时，重复上题，取 $n=2$.

9. 求出一个求积公式

$$\int_{-1}^1 f(x)\mathrm{d}x \approx c\sum_{i=0}^2 f(x_i)$$

使得这个公式对所有二次多项式是精确成立的.

10. a. 如果求积公式

$$\int_{-1}^1 f(x)\mathrm{d}x \approx f(\alpha) + f(-\alpha)$$

对所有二次多项式都精确成立，试问 α 应该取什么值? 对所有三次多项式，回答同样的问题.

对下列形式的多项式重复问题 a:

b. $f(x) = a + bx + cx^3 + dx^4$

c. $f(x) = a + \sum_{i=1}^n b_i x^{2i-1} + cx^{2n}$

11. 试问 α 取什么值的时候，公式

$$\int_0^2 f(x)\mathrm{d}x \approx f(\alpha) + f(2-\alpha)$$

在 Π_3 上精确成立?

12. 证明：若区间关于原点对称并且 w 是偶函数，则高斯结点也对称并且 $A_i = A_{n-i}$，$0 \leqslant i \leqslant n$.

13. 证明：每个下列形式的求积公式

$$\int_a^b f(x)\mathrm{d}x \approx \sum_{i=0}^n A_i f(x_i)$$

499

在 $C[a, b]$ 的某些有限维子空间上精确成立.

14. 证明：若

$$\int_a^b f(x)w(x)\mathrm{d}x = \sum_{i=0}^n A_i f(x_i)$$

对所有 $f \in \Pi_{2n+1}$ 成立，则多项式 $\prod_{i=0}^n (x - x_i)$ 在 $[a,b]$ 上与 Π_n 是关于 w 正交的.

15. 确定系数 A_0、A_1 和 A_2 使得公式

$$\int_0^2 f(x)\mathrm{d}x \approx A_0 f(0) + A_1 f(1) + A_2 f(2)$$

对于所有三次多项式精确成立.

16. 仅利用 $f(0)$, $f'(-1)$和 $f''(1)$, 计算 $\int_{-1}^{1} f(x)\mathrm{d}x$ 的一个近似公式使其对所有二次多项式都精确成立. 试问这个近似式对于三次多项式精确成立吗? 为什么?

17. 若公式

$$\int_0^1 xf(x)\mathrm{d}x = \sum_{i=0}^{4} A_i f(x_i)$$

对所有 $f \in \Pi_9$(九次多项式)是正确的, 则 x_0, x_1, …, x_4 一定是一个五次多项式 q 的零点, q 具有怎样的性质?

18. 若公式

$$\int_1^2 (x^4 - 1) f(x)\mathrm{d}x = Af(x_0) + Bf(x_1) + Cf(x_2)$$

对所有次数≤5 的多项式 f 都精确成立, 则 x_0, x_1 和 x_2 一定是具有怎样性质的多项式的根?

19. 下列哪个多项式在区间[0, 1]上关于权函数 $w(x)=1$ 与 Π_2 正交? $1+x$, $x-(1/2)$, x^2-3x+1, $35x^4-60x^3+32x-3$, x^3-3x^2+x-1.

20. 求出一非零多项式, 使它在区间[-1, 1]上关于权函数 $1+x^2$ 与 Π_2 正交.

21. 考虑下列形式的一个数值求积法则

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx Af\left(-\sqrt{\frac{3}{5}}\right) + Bf(0) + Cf\left(\sqrt{\frac{3}{5}}\right)$$

a. 在确定 A, B 和 C 所用的待定系数法中需要求解的线性方程组是什么? 并求解 A, B 和 C.

b. 为了确定牛顿-科茨公式中 A, B 和 C 需要计算的 3 个积分是什么? 求解 A, B 和 C.

22. 说明如何把高斯求积法则

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx \frac{5}{9} f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9} f(0) + \frac{5}{9} f\left(\sqrt{\frac{3}{5}}\right)$$

应用于 $\int_a^b f(x)\mathrm{d}x$. 用该结果计算:

a. $\int_0^{\frac{\pi}{2}} x\mathrm{d}x$

b. $\int_0^4 \frac{\sin t}{t}\mathrm{d}t$

500

23. 给定 $f(0)$, $f'(-1)$ 和 $f''(1)$, 利用待定系数法计算 $\int_{-1}^{1} x^2 f(x)\mathrm{d}x$ 的一个近似公式. 你给出的公式对所有 $f \in \Pi_2$ 应该给出精确的结果.

24. 求出 A, B 及 C 使得形如下式的数值求积法则

$$\int_{-1}^{1} xf(x)\mathrm{d}x \approx Af(-1) + Bf(0) + Cf(+1)$$

对所有最高次数是 m 的多项式精确成立. 试问 m 是多少?

25. 利用待定系数法, 求出下列法则中的 A, B 及 C, 它对于二次多项式应该给出精确的结果:

$$\int_{-3h}^{h} f(x)\mathrm{d}x \approx h[Af(0) + Bf(-h) + Cf(-2h)]$$

26. (续)如何能够利用上题中所给类型的一个数值求积法则, 得到一个用于求解下列常微分方程的法则?

$$\begin{cases} x' = f(t,x) \\ x(t_0) = x_0 \end{cases}$$

27. 求出下列高斯求积公式的系数和结点：

$$\int_0^1 x^4 f(x)\mathrm{d}x \approx A_0 f(x_0) + A_1 f(x_1)$$

28. 证明：不存在具有 n 个结点的高斯求积公式能在 Π_{2n} 上精确成立.
29. 如果函数 w 不是正的，那么在高斯求积理论中会发生什么情况？
30. 利用 7.2 节中的信息，求出关于区间$[-1, 1]$和权函数 $w(x)=\sqrt{1-x^2}$的高斯公式. 提示：利用习题 7.2.26～7.2.27.
31. 确定下列高斯公式的结点和权函数：

$$\int_{-1}^1 x^4 f(x)\mathrm{d}x \approx A_0 f(x_0) + A_1 f(x_1)$$

32. 推导出下列每个公式的结点和权函数的确切值. 并且确定使公式精确成立的最高次数多项式.
   a. (3)式
   b. (8)式
   c. (9)式
33. 利用下列公式确定 $\int_a^b f(x)\mathrm{d}x$ 的复合法则.
   a. 2 点高斯公式
   b. 3 点高斯公式
34. 如果我们取消对 q 次数的假设条件，试问高斯求积的定理 1 仍然成立吗？保留每一点 x_0，x_1，…，x_n 是 q 的零点的假设条件，但可能存在另外的零点.

计算机习题 7.3

编写一个计算机程序，对于所有函数 $p(x)(1-x^2)^{-1/2}$，其中 $p(x)\in\Pi_9$，验证 $n=4$ 的埃尔米特求积公式(4)式是否精确成立. 只需对 10 个检验函数 $T_k(x)(1-x^2)^{-1/2}$来验证公式就够了，其中 $T_k(x)=\cos(k\cos^{-1}x)$是 k 次切比雪夫多项式.

[501]

7.4 龙贝格积分

现在要介绍用龙贝格(Romberg)命名的一个算法，龙贝格首先给出了这种算法的递推形式. 假设需要积分

$$I=\int_a^b f(x)\mathrm{d}x \tag{1}$$

的近似值. 在讨论过程中函数 $f(x)$和区间$[a, b]$将保持不变.

7.4.1 递推梯形法则

设 $T(n)$表示在长度是 $h=(b-a)/n$ 的 n 个子区间上积分 I 的梯形法则. 根据 7.2 节中(4)式，我们有

$$T(n)=h\sum_{i=0}^{n}{}''f(a+ih)=\frac{(b-a)}{n}\sum_{i=0}^{n}{}''f\left(a+i\frac{(b-a)}{n}\right) \tag{2}$$

这里求和符号中的两撇表示和式中第一项和最后一项减半.

例 1 当区间是$[0, 1]$时，$T(1)$，$T(2)$，$T(4)$和 $T(8)$的显式公式是什么？

解 利用(2)式，我们有

$$T(1)=\frac{1}{2}f(0)+\frac{1}{2}f(1)$$

$$T(2)=\frac{1}{4}f(0)+\frac{1}{2}\left[f\left(\frac{1}{2}\right)\right]+\frac{1}{4}f(1)$$

$$T(4)=\frac{1}{8}f(0)+\frac{1}{4}\left[f\left(\frac{1}{4}\right)+f\left(\frac{1}{2}\right)+f\left(\frac{3}{4}\right)\right]+\frac{1}{8}f(1)$$

$$T(8)=\frac{1}{16}f(0)+\frac{1}{8}\left[f\left(\frac{1}{8}\right)+f\left(\frac{1}{4}\right)+f\left(\frac{3}{8}\right)+f\left(\frac{1}{2}\right)+f\left(\frac{5}{8}\right)+f\left(\frac{3}{4}\right)+f\left(\frac{7}{8}\right)\right]+\frac{1}{16}f(1)$$

■

可以看出，如果要计算 $T(2n)$，则可以利用 $T(n)$的计算中已有的结果，只需要计算那些出现在 $T(2n)$中而没有出现在 $T(n)$中的项. 例如，根据上面的公式，我们可以看出

$$T(2)=\frac{1}{2}T(1)+\frac{1}{2}\left[f\left(\frac{1}{2}\right)\right]$$

$$T(4)=\frac{1}{2}T(2)+\frac{1}{4}\left[f\left(\frac{1}{4}\right)+f\left(\frac{3}{4}\right)\right]$$

$$T(8)=\frac{1}{2}T(4)+\frac{1}{8}\left[f\left(\frac{1}{8}\right)+f\left(\frac{3}{8}\right)+f\left(\frac{5}{8}\right)+f\left(\frac{7}{8}\right)\right]$$

取 $h=(b-a)/2n$，关于任意区间$[a,b]$上的一般公式如下：

$$T(2n)=\frac{1}{2}T(n)+h[f(a+h)+f(a+3h)+f(a+5h)+\cdots+f(a+(2n-1)h)]$$

502

或者

$$T(2n)=\frac{1}{2}T(n)+h\sum_{i=1}^{n}f(a+(2i-1)h) \tag{3}$$

为了证明等式(3)，首先利用 $T(n)$表示 $T(2n)$：

$$T(2n)=\frac{1}{2}T(n)+\left[T(2n)-\frac{1}{2}T(n)\right]$$

利用(2)式，可重写括号中的表达式为

$$T(2n)-\frac{1}{2}T(n)=h\sum_{i=0}^{2n}{}''f(a+ih)-h\sum_{i=0}^{n}{}''f(a+2ih)=h\sum_{i=1}^{n}f(a+(2i-1)h)$$

因为第二个和式中的各项形如 $f(a+2ih)$，所以它们与第一个和式中所有偶数项相抵消. 在所得和式中，i 的范围是这样确定的：第一个和式中的第一项奇数项是 $f(a+ih)$，所以我们取 $2i-1=1$；最后的奇数项是 $f(a+(2n-1)h)$，所以我们取 $2i-1=2n-1$. 因此，等式(3)成立. 这样(1)式中积分 I 可以用递推梯形法则近似计算. 如果有 2^n 个相同的子区间，则(3)式给出递推梯形法则：

$$T(2^n)=\frac{1}{2}T(2^{n-1})+h_n\sum_{i=1}^{2^{n-1}}f(a+(2i-1)h_n)$$

其中

$$h_0=b-a \quad h_n=h_{n-1}/2 \quad (n\geqslant 1)$$

7.4.2 龙贝格算法

在龙贝格算法中使用上述公式. 设 $R(n, 0)$表示具有 2^n 个子区间的梯形估计，我们有

$$\begin{cases} R(0,0) = \frac{1}{2}(b-a)[f(a)+f(b)] \\ R(n,0) = \frac{1}{2}R(n-1,0) + h_n \sum_{i=1}^{2^{n-1}} f(a+(2i-1)h_n) \end{cases} \tag{4}$$

对于一个适度的 M 值，计算 $R(0, 0)$，$R(1, 0)$，$R(2, 0)$，…，$R(M, 0)$，并且其中没有重复的函数值的计算. 在龙贝格算法的其余部分中，还要计算附加值 $R(n, m)$. 所有这些都可以被理解为积分 I 的估计. 计算出 $R(M, 0)$以后，不再需要被积函数 f 值的计算. 根据公式

503

$$R(n,m) = R(n,m-1) + \frac{1}{4^m-1}[R(n,m-1) - R(n-1,m-1)] \tag{5}$$

对于 $n\geqslant 1$ 和 $m\geqslant 1$ 构造 R 阵列的各列. 这一计算是非常简单的，最终可以得到下面形式的阵列

$$\begin{array}{lllllll} R(0,0) & & & & & & \\ R(1,0) & R(1,1) & & & & & \\ R(2,0) & R(2,1) & R(2,2) & & & & \\ R(3,0) & R(3,1) & R(3,2) & R(3,3) & & & \\ R(4,0) & R(4,1) & R(4,2) & R(4,3) & R(4,4) & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \\ R(M,0) & R(M,1) & R(M,2) & R(M,3) & R(M,4) & \cdots & R(M,M) \end{array}$$

下面是按行计算的龙贝格算法的伪代码：

```
input a, b, M
h ← b − a
R(0, 0) ← ½(b−a)[f(a)+f(b)]
for n = 1 to M do
    h ← h/2
    R(n, 0) ← ½R(n−1, 0) + h Σ_{i=1}^{2^{n−1}} f(a+(2i−1)h)
    for m = 1 to n do
        R(n, m) ← R(n, m−1) + [R(n, m−1) − R(n−1, m−1)]/(4^m − 1)
    end do
end do
output R(n, m)   (0≤n≤M, 0≤m≤n)
```

因为需要计算 2^M+1 个函数值，所以通常只选取一个适度的 M 值. 更为精致的算法应该包含一个自动终止程序，当达到指定的误差标准时停止计算.

7.4.3 分析

为了说明(5)式的来源，我们从**欧拉-麦克劳林公式**入手：

$$\int_0^1 f(t)\,\mathrm{d}t = \frac{1}{2}[f(0)+f(1)]$$

$$+\sum_{k=1}^{m-1}A_{2k}[f^{(2k-1)}(0)-f^{(2k-1)}(1)] \tag{6}$$
$$-A_{2m}f^{(2m)}(\xi_0)$$

其中 ξ_0 是 0 与 1 之间的一点．可以证明对于任意函数 $f\in C^{2m}[0,1]$，(6)式都成立．7.7 节中给出了这个重要公式的证明．常数 $k!A_k$ 称为**伯努利数**．它们可以由下列等式定义 [504]

$$\frac{x}{e^x-1}=\sum_{k=0}^{\infty}A_kx^k$$

定义函数 $g(t)=f(x_i+ht)$，其中 $h=x_{i+1}-x_i$，把(6)式应用于 g，并且在得到的积分中作变量替换 $t=(x-x_i)/h$，由基本公式(6)可得：

$$\int_{x_i}^{x_{i+1}}f(x)\mathrm{d}x=\frac{h}{2}[f(x_i)+f(x_{i+1})] \tag{7}$$
$$+\sum_{k=1}^{m-1}A_{2k}h^{2k}[f^{(2k-1)}(x_i)-f^{(2k-1)}(x_{i+1})]$$
$$-A_{2m}h^{2m+1}f^{(2m)}(\xi_i)$$

在(7)式两端作运算 $\sum_{i=0}^{2^n-1}$．如果 $x_i=a+ih, 0\leqslant i\leqslant 2^n, h=(b-a)/2^n$，则

$$\int_a^b f(x)\mathrm{d}x=\frac{h}{2}\sum_{i=0}^{2^n-1}[f(x_i)+f(x_{i+1})] \tag{8}$$
$$+\sum_{k=1}^{m-1}A_{2k}h^{2k}[f^{(2k-1)}(a)-f^{(2k-1)}(b)]$$
$$-A_{2m}(b-a)h^{2m}f^{(2m)}(\xi)$$

其中 $\xi\in(a,b)$．这个等式中的误差项如习题 7.4.2 所述的一样处理．(8)式右端的第一项恰好是(1)式中积分 I 在长度为 $h=(b-a)/2^n$ 的子区间上的梯形估计．从而，由(8)式可知

$$I=R(n,0)+c_2h^2+c_4h^4+c_6h^6+\cdots+c_{2m-2}h^{2m-2}+c_{2m}h^{2m}f^{(2m)}(\xi) \tag{9}$$

其中 $f\in C^{2m}[a,b]$，$\xi\in(a,b)$．对于任意的 h，该等式都成立，而且系数 $c_2, c_4, \cdots, c_{2m}$ 与 h 无关．因为(9)式是 7.1 节中(20)式的特殊情形，所以根据 7.1 节中的理查森外推分析，直接可知(4)式成立．

为了充分利用龙贝格算法，我们需要函数 f 属于类 $C^{2m}[a,b]$，并选取尽可能大的 m 值．这个假设保证了可以使用欧拉-麦克劳林公式，并且表明量值 $R(n,m)$ 收敛于 f 的积分，具有误差 $\mathcal{O}(h^{2m})$，其中 $h=(b-a)2^{-n}$．如果仅仅假设 f 是连续的会是什么结果？

定理 1（龙贝格算法收敛性定理） 若 $f\in C[a,b]$，则龙贝格阵列中每一列都收敛于 f 的积分．因此，对每个 m，

$$\lim_{n\to\infty}R(n,m)=\int_a^b f(x)\mathrm{d}x$$ [505]

证明 我们从第一列开始，它包含积分 I 的梯形估计．具有 k 个子区间的梯形法则可以写成

$$h\sum_{i=0}^{k}{}''f(a+ih)=\frac{1}{2}h\sum_{i=0}^{k-1}f(a+ih)+\frac{1}{2}h\sum_{i=1}^{k}f(a+ih)$$

等式的右端是 I 的两个黎曼和的平均值．因为 $h=(b-a)/k$，所以当 $k\to\infty$时子区间的最大长度趋向于 0．因此，根据黎曼积分理论，两个黎曼和都收敛于 I．它们的平均值当然也收敛于 I．这就证明了 $\lim\limits_{n\to\infty}R(n,0)=I$．对于第二列，我们注意到

$$R(n,1)=\frac{4}{3}R(n,0)-\frac{1}{3}R(n-1,0)$$

由此可得

$$\lim_{n\to\infty}R(n,1)=\frac{4}{3}I-\frac{1}{3}I=I$$

用同样的方法可对接下去的各列进行分析. ■

习题 7.4

1. 由(6)式推导出(7)式.
2. 由(7)式推导出(8)式，特别地，证明

$$从\ h^{2m+1}\sum_{i=0}^{2^n-1}f^{(2m)}(\xi_i)\ 可转换成(b-a)h^{2m}f^{(2m)}(\xi)$$

3. 取 $h=1/2^n$，建立下面的等式

$$I=\frac{4}{3}T\left(f,\frac{h}{2}\right)-\frac{1}{3}T(f,h)-\sum_{n=1}^{\infty}\frac{4^n-1}{3(4^n)}c_{2n+2}h^{2n+2}$$

其中

$$I=\int_0^1 f(x)\mathrm{d}x, T(f,h)=h\sum_{i=0}^{2^n}{}'' f(ih)$$

4. 证明：龙贝格阵列中第二列是对 f 应用辛普森法则的结果．(见 7.2 节中(6)式.)
5. 用数学归纳法证明

$$I-R(n,m-1)=ah^{2m}+bh^{2m+2}+ch^{2m+4}+\cdots$$

6. 应用龙贝格算法求出下列积分的 $R(2,2)$：

   a. $\int_1^3\frac{\mathrm{d}x}{x}$

   b. $\int_0^{\pi/2}\left(\frac{x}{\pi}\right)^2\mathrm{d}x$ (用含 π 的项)

7. 假设 $S(f,h)$是(1)式中积分 I 的求积法则并且误差级数是 $c_4h^4+c_6h^6+\cdots$．把 $S(f,h)$和 $S(f,h/3)$结合起来，求出 I 的一个更精确的近似.

[506]

8. 在龙贝格算法中，$R(n,0)$是利用 2^n 个子空间上的梯形法则对 $\int_a^b f(x)\mathrm{d}x$ 的估计．对于 $0\leqslant i\leqslant N$ 和 $0\leqslant j\leqslant N$，试问计算 $R(i,j)$需要多少次 $f(x)$的赋值?
9. 如果梯形法则满足等式

$$\int_a^b f(x)\mathrm{d}x=T(f,h)+c_1h+c_2h^2+c_3h^2+\cdots$$

而不是满足(9)式，那么将如何修正公式(5)?

10. 在龙贝格算法中，第二列元素满足

$$R(i,1)=I+C_4h_i^4+C_6h_i^6+\cdots$$

其中 $I=\int_a^b f(x)\mathrm{d}x$，$h_i=(b-a)/2^i$．推导出计算第三列元素和它的误差级数中第一项的公式.

11. (**Milne 法则**)用龙贝格阵列中第一列元素表示 $R(2,2)$. 证明 $R(3,3)$不是牛顿–科茨公式而 $R(2,2)$是这样的公式.
12. 直接根据下列事实

$$\sum_{0\leqslant i\leqslant 2n} f(a+ih) - \sum_{\substack{0\leqslant i\leqslant 2n\\ i\text{为偶数}}} f(a+ih) = \sum_{\substack{0\leqslant i\leqslant 2n\\ i\text{为奇数}}} f(a+ih)$$

证明等式(3)成立.

计算机习题 7.4

编写一个子程序，用来执行定义在任意区间$[a, b]$上的函数 f 的龙贝格算法. 用户要具体指定阵列中所计算的行数，并且当计算完成后要看到整个阵列. 编写一个主程序并且用下列积分测试你的龙贝格子程序：

a. $\int_0^1 \frac{\sin x}{x}\mathrm{d}x$

b. $\int_{-1}^1 \frac{\cos x - e^x}{\sin x}\mathrm{d}x$

c. $\int_1^\infty (xe^x)^{-1}\mathrm{d}x$

编写这些积分的程序，要避免由于减法而产生有效数字的严重丢失. 习惯上也用等式 $f(x_0)=\lim_{x\to x_0} f(x)$定义任何可疑点 x_0 上的函数 f. 如果极限存在，则这种方法便保证了 f 在 x_0 点的连续性. 对于第三个积分，作适当的变量替换，例如 $x=1/t$. 计算出龙贝格阵列中的 7 行. 打印出每一种情形的阵列，并要求打印格式能够反映出收敛性.

7.5 自适应求积

自适应求积方法是指自动地利用被积函数的性质来计算定积分，理想情况下，为了计算积分

507

$$\int_a^b f(x)\mathrm{d}x \tag{1}$$

用户只需提供被积函数 f、区间$[a, b]$以及所要求的精度 ε. 然后程序将区间划分为各种长度的小段使得在子区间上数值积分产生满足精度要求的结果.

这里讨论一个典型的自适应求积方法，主要用到子区间上的辛普森法则连同所涉及的误差估计. 7.2 节中(6)式给出**辛普森法则**：

$$\int_a^b f(x)\mathrm{d}x = S(a,b) - \frac{1}{90}[(b-a)/2]^5 f^{(4)}(\xi)$$

$$S(a,b) = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] \tag{2}$$

其中 $\xi\in(a, b)$. 其主要思想是，如果在一给定子区间上的辛普森法则不满足精度要求，则该区间将被等分为两部分，在每个长度减半的区间上应用辛普森法则，重复这个过程得到积分的一个近似，它在所有涉及的子区间上具有相同的精度. 最后，我们计算应用 n 次辛普森法则的积分，

$$\int_a^b f(x)\mathrm{d}x = \sum_{i=1}^n \int_{x_{i-1}}^{x_i} f(x)\mathrm{d}x = \sum_{i=1}^n (S_i + e_i) = \sum_{i=1}^n S_i + \sum_{i=1}^n e_i$$

其中 S_i 是区间$[x_{i-1}, x_i]$上积分的近似，e_i 是相应的局部误差. 如果

$$|e_i| \leqslant \varepsilon(x_i - x_{i-1})/(b-a) \tag{3}$$

则整体误差界为

$$\left|\sum_{i=1}^{n} e_i\right| \leqslant \sum_{i=1}^{n} |e_i| \leqslant \frac{\varepsilon}{b-a}\sum_{i=1}^{n}(x_i - x_{i-1}) = \varepsilon$$

从而，局部误差准则(3)导出绝对误差界

$$\left|\int_a^b f(x)\mathrm{d}x - \sum_{i=1}^{n} S_i\right| \leqslant \varepsilon$$

根据(2)式，下式给出区间$[u, v]$上的基本辛普森法则：

508

$$\int_u^v f(x)\mathrm{d}x = S(u,v) - \frac{1}{90}[(v-u)/2]^5 f^{(4)}(\xi_1) \tag{4}$$

其中 $\xi_1 \in (u, v)$. 如果求积区间在中点 $w=(u+v)/2$ 处等分为两个子区间，那么对每个子区间利用辛普森法则，可计算出更精确的积分值. 这样做的结果如下：

$$\begin{aligned}\int_u^v f(x)\mathrm{d}x &= \int_u^w f(x)\mathrm{d}x + \int_w^v f(x)\mathrm{d}x \\ &= S(u,w) - \frac{1}{90}[(w-u)/2]^5 f^{(4)}(\xi_2) + S(w,v) \\ &\quad - \frac{1}{90}[(v-w)/2]^5 f^{(4)}(\xi_3) \\ &= S^* + S^{**} - \frac{1}{90}\left(\frac{v-u}{2^2}\right)^5[f^{(4)}(\xi_2) + f^{(4)}(\xi_3)] \\ &= S^* + S^{**} - \frac{1}{2^9}\cdot\frac{1}{90}(v-u)^5 f^{(4)}(\xi)\end{aligned} \tag{5}$$

在这个计算过程中，我们已经规定

$$S^* \equiv S(u,w) \qquad S^{**} \equiv S(w,v)$$

$$f^{(4)}(\xi) \equiv \frac{1}{2}[f^{(4)}(\xi_2) + f^{(4)}(\xi_3)] \tag{6}$$

其中 $\xi_2 \in (u, w)$，$\xi_3 \in (w, v)$，$\xi \in (u, v)$. 如果假设 $f^{(4)}$ 连续，则(6)式成立. 像通常的这种公式一样，不能估计误差项，也不能确定它的界，除非 $f^{(4)}$ 的某些信息是可利用的. 然而，对于自动计算过程来说，有必要找出一种估计 $f^{(4)}(\xi)$大小的方法. 通过假设在所有小区间上 $f^{(4)}$ 是常数，我们可以实现这个目标. 特别地，在(4)式和(5)式中我们假设 $f^{(4)}(\xi_1) = f^{(4)}(\xi)$，则(5)式乘以 16/15 减去(4)式乘以 1/15，便可以消去含有 $f^{(4)}$ 的项，其结果是

$$\int_u^v f(x)\mathrm{d}x \approx S^* + S^{**} + \frac{1}{15}[S^* + S^{**} - S(u,v)] \tag{7}$$

我们已经对自适应求积方法的基本思路作了铺垫. 假设就给定容许误差 ε 来说，可以数值计算积分(1). 如果 $f^{(4)}$ 存在并且缓慢地变化，则在小区间上(7)式中的近似将是令人满意的. 自适应算法从区间$[a, b]$开始考虑，对这个区间(与随后考虑的其他区间一样)构造具有 6 个分量的向量：

$$[a,h,f(a),f(a+h),f(a+2h),S] \quad h=\frac{1}{2}(b-a) \tag{8}$$

该向量包含辛普森估计 $S=S(a,\ b)$，以及用等式

509

$$S=\frac{h}{3}[f(a)+4f(a+h)+f(a+2h)]$$

计算它所需要的数据. 接着计算 $c=a+h$，$S^{*}=S(a,\ c)$及 $S^{**}=S(c,\ b)$. 如前面解释的那样，$S^{*}+S^{**}$是积分更精确的估计. 为了观察它是否足够好(即是否满足容许误差 ε)，我们将利用(7)式和(3)式. 检验不等式

$$|S^{*}+S^{**}-S|/15<\varepsilon(2h)/(b-a) \tag{9}$$

如果不等式成立，则依照(7)式，$S^{*}+S^{**}+[S^{*}+S^{**}-S]/15$ 可以作为积分(1)的值. 如果不等式(9)不成立，则把区间划分为两个等长的子区间；即$[a,\ c]$和$[c,\ b]$，其中 $c=(a+b)/2$. 对每个小区间，我们像前面那样构造向量. 丢弃(8)式中的向量，并且在它的位置上有两个新向量：

$$[a,h/2,f(a),f(y),f(c),S^{*}] \quad y=a+h/2$$
$$[c,h/2,f(c),f(z),f(b),S^{**}] \quad z=c+h/2$$

注意，在计算 $S^{*}=S(a,\ c)$和 $S^{**}=S(c,\ b)$时，由于所有其他数据已经被算出，并且已算出的数据用特殊的格式加以存储，所以只需要两个新的函数 $f(y)$和 $f(z)$赋值. 应用于(8)式中向量的过程最后将应用于算法生成的每一个向量.

以下描述算法中的一般步骤：在任一步，在某些子区间上 f 的积分的和将累加到变量Σ中. 同时，存在早先所述的向量的工作栈，每个向量对应一个区间，并且该区间上 f 的积分还未满意算出，这些向量之一，例如

$$[u,h,\bar{u},\bar{w},\bar{v},S]$$

取自工作栈. 它具有下列性质：

$$w=u+h \quad v=u+2h$$
$$\bar{u}=f(u) \quad \bar{w}=f(w) \quad \bar{v}=f(v)$$
$$S=(\bar{u}+4\bar{w}+\bar{v})h/3$$

其次，用 $h/2$ 替换 h，计算得到

$$y=u+h \quad z=u+3h$$
$$\bar{y}=f(y) \quad \bar{z}=f(z)$$
$$S^{*}=(\bar{u}+4\bar{y}+\bar{w})h/3$$
$$S^{**}=(\bar{w}+4\bar{z}+\bar{v})h/3$$

作为$\int_u^v f(x)\mathrm{d}x$ 的估计值，我们希望知道 $S^{*}+S^{**}$是否能通过误差检验，其检验条件是

$$|S^{*}+S^{**}-S|\leqslant 60\varepsilon h/(b-a) \tag{10}$$

510

如果通过检验，则把 $S^{*}+S^{**}+(S^{*}+S^{**}-S)/15$ 增加至Σ中，并且从工作栈中选取一个新的向量继续该过程. 如果检验失败，则把两个新的向量添加到工作栈中：

$$[u,h,\bar{u},\bar{y},\bar{w},S^{*}]$$
$$[u+2h,h,\bar{w},\bar{z},\bar{v},S^{**}]$$

在该步骤中，从工作栈中去掉了另外一个向量并且像前面那样继续该过程. 工作栈中的向量不能超过 n 个，n 是用户设置的参数. 这有助于防止算法不能终止.

下面给出这个算法的伪代码．它来自刚才给出的描述过程．要关注那些后面可能用到的量值的存储．工作栈中的向量记为 $v^{(1)}$，$v^{(2)}$，等等．每个向量有 6 个分量：

$$v^{(k)}=[v_1^{(k)},v_2^{(k)},\cdots,v_6^{(k)}]$$

第一个分量 $v_1^{(k)}$ 始终表示区间的左端点；$v_2^{(k)}$ 是这个区间长度的一半；$v_3^{(k)}$，$v_4^{(k)}$ 及 $v_5^{(k)}$ 分别是 f 在区间左端点、中点及右端点上的值；$v_6^{(k)}$ 是这个区间上由辛普森法则给出的值．

完整的辛普森自适应求积算法如下：

```
input a, b, ε, n
Δ ← b − a;  Σ ← 0;  h ← Δ/2;  c ← (a + b)/2;  k ← 1
ā ← f(a);  b̄ ← f(b);  c̄ ← f(c)
S ← (ā + 4c̄ + b̄)h/3
v^(1) ← [a, h, ā, c̄, b̄, S]
while 1 ≤ k ≤ n
    h ← v_2^(k)/2
    ȳ ← f(v_1^(k) + h)
    S* ← (v_3^(k) + 4ȳ + v_4^(k))h/3
    z̄ ← f(v_1^(k) + 3h)
    S** ← (v_4^(k) + 4z̄ + v_5^(k))h/3
    if |S* + S** − v_6^(k)| < 60εh/Δ then
            Σ ← Σ + S* + S** + [S* + S** − v_6^(k)]/15
            k ← k − 1
            if k ≤ 0 then output Σ; exit
    else
            if k ≥ n then output failure; exit
               v̄ ← v_5^(k)
            v^(k) ← [v_1^(k), h, v_3^(k), ȳ, v_4^(k), S*]
            k ← k + 1
            v^(k) ← [v_1^(k−1) + 2h, h, v_5^(k−1), z̄, v̄, S**]
    end if
end while
```

511

在一些编程语言中，重要的是把最频繁存取的量存放在阵列的同一列中，使其对应于相邻存储位置．如果是在这种情况下，则工作栈向量可储存在 2 维数组 V(I，K)$\leftarrow v_i^{(k)}$ 中．

这里我们介绍了具有显式工作栈的自适应求积算法，但它本质上是一个递归过程．首先，试用简单的辛普森法则，如果误差满足要求，那么就认可答案；否则，则把区间等分为两部分并且在每个小区间上递归地调用上述过程．利用递归性质，我们更容易理解算法的概念．此外，由于大多数编程语言都支持递归关系，所以编写递归算法的代码是简单易行的．我们把它留作计算机习题 7.5.1．

习题 7.5

1．用局部误差标准（（3）式）得到绝对误差界．试建立局部误差标准并由此导出相对误差界

$$\left|\int_a^b f(x)\,dx-\sum_{i=1}^{n}S_i\right|\leqslant\left|\int_a^b f(x)\,dx\right|\varepsilon$$

2. 利用梯形法则

$$\int_u^v f(x)\mathrm{d}v = T(u,v) - \frac{1}{12}(v-u)^3 f''(\xi)$$

其中

$$T(u,v) = \frac{1}{2}(v-u)[f(u)+f(v)]$$

建立类似于(7)式的一个近似公式.

3. 复合梯形法则可写为

$$I = T_m - \frac{1}{12}(v-u)h^2 f''(\xi)$$

其中

$$I = \int_u^v f(x)\mathrm{d}x$$

$$T_m = \frac{h}{2}[f(u) + 2\sum_{i=1}^{m-1} f(u+ih) + f(v)]$$

试说明怎样结合 T_m 和 T_{2m} 来得到 I 的一个更好的近似. 这里 m 是区间$[u, v]$中子区间的个数.

4. 复合辛普森法则可写为

$$I = S_{2m} - \frac{1}{180}(v-u)h^4 f^{(4)}(\xi)$$

其中

$$I = \int_u^v f(x)\mathrm{d}x$$

$$S_{2m} = \frac{h}{3}[f(u) + 2\sum_{i=2}^{m} f(x_{2i-2}) + 4\sum_{i=1}^{m} f(x_{2i-1}) + f(v)]$$

512

并且 $h=(v-u)/(2m)$, $x_i=u+ih(0\leqslant i\leqslant 2m)$. 这里区间$[u, v]$中有 $2m$ 个子区间. 试说明怎样结合 S_{2m} 和 S_{4m} 来得到 I 的一个更好的近似.

5. 在 7.4 节中, 在$[u, v]$上具有 m 个子区间的梯形法则与$[u, v]$上具有 $2m$ 个子区间的梯形法则之间, 有下列关系存在

$$T_{2m} = \frac{1}{2}T_m + h\sum_{i=1}^{m} f(x_{2i-1})$$

其中 $h=(v-u)/(2m)$, $x_i=u+ih(0\leqslant i\leqslant 2m)$. 试问它可以用来建立一个自适应机制吗?

计算机习题 7.5

1. 编写自适应算法程序并对下列积分测试这个程序:

a. $\int_0^1 x^{1/2}\mathrm{d}x$

b. $\int_0^1 (1-x)^{1/2}\mathrm{d}x$

c. $\int_0^1 (1-x)^{1/4}\mathrm{d}x$

2. 编写并测试递归过程形式的自适应算法程序.

7.6 逼近泛函的 Sard 定理

线性空间上的一个**线性泛函**是线性空间到纯量域(本书中常用$\mathbb{R}$)的一个线性映射. 例如,如果线性空间是 $C[a, b]$, 一个重要的线性泛函 φ 定义为

$$\varphi(f) = \int_a^b f(x)\mathrm{d}x \quad f \in C[a,b]$$

在数值计算问题中,最基本的泛函是**点赋值**. 在$[a, b]$中取定 x, 下列等式定义一个线性泛函 $\hat{x}$

$$\hat{x}(f) = f(x) \quad f \in C[a,b]$$

利用点赋值的线性组合,我们得到用途更广的泛函 ψ

$$\psi = \sum_{i=0}^{n} c_i \hat{x}_i \quad \text{即} \quad \psi(f) = \sum_{i=0}^{n} c_i f(x_i) \tag{1}$$

513 人们有理由认为,在数值计算中,这是可直接计算的泛函中最一般的类型;其他泛函(如积分)一定会被 ψ 这样的逼近泛函替换. 这正是数值积分和数值微分问题中所要做的.

Arthur Sard 在 1940~1970 年发展了逼近泛函的相关理论. 它用一种有趣的方式与自然样条相关联. 我们要逼近的这类泛函由下列等式给出

$$\varphi(f) = \sum_{i=0}^{N} \left\{ \int_a^b \alpha_i(x) f^{(i)}(x)\mathrm{d}x + \sum_{j=1}^{n} \beta_{ij} f^{(i)}(z_{ij}) \right\} \tag{2}$$

点 z_{ij} 在$[a, b]$中,假定函数 α_i 在$[a, b]$上分段连续. 泛函 φ 可作用于任一 $f \in C^N[a, b]$.

如果对于每个 $f \in W$ 都有 $\varphi(f)=0$, 则称泛函 φ **零化**空间 W. 像(2)式中那样的泛函的**佩亚诺核**是函数

$$K_m(t) = \frac{1}{m!}\varphi_x[(x-t)_+^m]$$

其中 $m \geqslant N$ 并且

$$(x-t)_+^m = \begin{cases} (x-t)^m & x \geqslant t \\ 0 & x < t \end{cases}$$

符号 φ_x 表示泛函作为 x 的函数作用于$(x-t)_+^m$. 当 $t=0$ 时得到的函数 x_+^m 称为**截断幂函数**.

例 1 考虑下列等式定义的泛函 φ

$$\varphi(f) = \int_0^\pi (\cos x) f'(x)\mathrm{d}x$$

这是(2)式在 $N=1$ 及 $\alpha_1(x)=\cos x$ 时的情形. 试问什么是 φ 的佩亚诺核 K_1?

解 可用佩亚诺核的定义直接计算. 注意到

$$\frac{\mathrm{d}}{\mathrm{d}x}(x-t)_+^m = m(x-t)_+^{m-1} \quad (m \geqslant 1)$$

因而,在这个例题中,

$$K_1(t) = \varphi_x[(x-t)_+^1] = \int_0^\pi (\cos x) \frac{\mathrm{d}}{\mathrm{d}x}(x-t)_+ \,\mathrm{d}x$$

$$= \int_0^\pi (\cos x)(x-t)_+^0 \,\mathrm{d}x$$

$$= \int_t^\pi (\cos x)\mathrm{d}x = -\sin t$$ ■ [514]

1905 年证明了下面的结果.（其中 Π_m 表示次数最多是 m 次的全体多项式的空间.）

定理 1(佩亚诺核定理) 若(2)式中的泛函零化 Π_m，则对所有 $f \in C^{m+1}[a, b]$，

$$\varphi(f) = \int_a^b K_m(t) f^{(m+1)}(t)\mathrm{d}t \tag{3}$$

其中 $m \geqslant N$ 并且 $K_m(t)$ 是 φ 的佩亚诺核.

证明 根据带有积分余项的泰勒定理(1.1 节中定理 5)，

$$f(x) = \sum_{k=0}^{m} \frac{1}{k!} f^{(k)}(a)(x-a)^k + r(x)$$

$$r(x) = \frac{1}{m!}\int_a^x f^{(m+1)}(t)(x-t)^m \mathrm{d}t$$

因为 φ 零化 Π_m，所以 $\varphi(f) = \varphi(r)$. 把 r 表示为

$$r(x) = \frac{1}{m!}\int_a^b f^{(m+1)}(t)(x-t)_+^m \,\mathrm{d}t$$

由此可知

$$\varphi(r) = \frac{1}{m!}\int_a^b f^{(m+1)}(t)\varphi_x[(x-t)_+^m]\mathrm{d}t$$

这一步包括把泛函 φ 转移到积分符号里面，可根据微分中的某些定理证明这一点成立，它需要用到我们对 φ 采用的假设条件. ■

例 2 Sard[1963，第 31 页]给出的实例是泛函

$$\varphi(f) = \int_0^1 f(x) x^{-1/2}\mathrm{d}x$$

求一个逼近公式

$$\psi(f) = c_1 f(0) + c_2 f(1)$$

它对于多项式空间 Π_1 上的 φ 恰好是代入过程. 再求出这个公式的误差.

解 我们要求 $\varphi - \psi$ 零化 Π_1. 利用待定系数法(如 7.2 节中)，首先取 $f(x)=1$，得

[515]

$$0 = \varphi(f) - \psi(f) = \int_0^1 x^{-1/2}\mathrm{d}x - (c_1 + c_2) = 2 - c_1 - c_2$$

取 $f(x)=x$，得

$$0 = \varphi(f) - \psi(f) = \int_0^1 x^{1/2}\mathrm{d}x - c_2 = \frac{2}{3} - c_2$$

因此，$c_2 = 2/3$，$c_1 = 4/3$. 泛函 $\varphi - \psi$ 的佩亚诺核是

$$\begin{aligned}(\varphi_x - \psi_x)(x-t)_+^1 &= \int_0^1 (x-t)_+ \, x^{-1/2}\mathrm{d}x - \frac{4}{3}(0-t)_+ - \frac{2}{3}(1-t)_+ \\ &= \int_t^1 (x-t) x^{-1/2}\mathrm{d}x - \frac{2}{3}(1-t) \\ &= \frac{4}{3} t (t^{1/2} - 1)\end{aligned}$$

佩亚诺核定理应用于 $\varphi-\psi$，得到

$$\int_0^1 f(x)x^{-1/2}\mathrm{d}x-\left[\frac{4}{3}f(0)+\frac{2}{3}f(1)\right]=\int_0^1\frac{4}{3}t(t^{1/2}-1)f''(t)\mathrm{d}t$$

当 $f\in C^2[0,1]$时，该等式的右边恰好表示误差．因为在区间[0，1]上核是非正的，所以由积分中值定理(1.1节)知

$$\int_0^1\frac{4}{3}t(t^{1/2}-1)f''(t)\mathrm{d}t=f''(\xi)\int_0^1\frac{4}{3}t(t^{1/2}-1)\mathrm{d}t=-\frac{2}{15}f''(\xi)$$ ■

如果 φ 和 ψ 是 Π_m 上相同的两个泛函，则它们的差零化 Π_m 并且有佩亚诺核 K_m，再由柯西-施瓦茨不等式知

$$|\varphi(f)-\psi(f)|\leqslant\|K_m\|_2\|f^{(m+1)}\|_2 \tag{4}$$

下标是2的范数是指$[a,b]$上常用的 L_2 范数．如果 ψ 中的某些参数不能完全由 $\varphi-\psi$ 零化 Π_m 来确定的话，则可以通过使因子 $\int_a^b K_m(t)^2\mathrm{d}t$ 极小化来选取这些参数．用这种方式，我们得到一个泛函 ψ，它是 Sard 意义下 φ 的一个最佳逼近．Schoenberg 发现可以用更简单的方法得到这种最佳公式，下面给出其重要的定理．在定理中，$\varphi\circ L$ 是 φ 和 L 的合成，即$(\varphi\circ L)(f)=\varphi(Lf)$.

定理2(Schoenberg 定理)　*设 φ 是(2)式给出的线性泛函．给定结点：$a=t_0<t_1<\cdots<t_n=b$，其中 $n>N$．在所有形如 $\sum_{i=0}^{n}c_i\hat{t}_i$ 并且在 Π_m 上与 φ 一致的泛函中，Sard 意义下 φ 的最佳逼近是 $\varphi\circ L$，其中 L 是产生给定结点上 $2m+1$ 次自然样条插值的线性算子．*

516

证明　设 ψ 具有形式 $\psi=\sum_{i=0}^{n}c_i\hat{t}_i$，并且假设对任一 $p\in\Pi_m$，有 $\psi(p)=\varphi(p)$，那么 $\varphi-\psi$ 零化 Π_m．设 K_m 是它的佩亚诺核．如果 f 是给定结点上的 $2m+1$ 次自然样条，那么 $Lf=f$．因为次数$\leqslant m$ 的多项式也是这种自然样条，所以对 $p\in\Pi_m$ 都有 $Lp=p$．因此，$\varphi-\varphi\circ L$ 也零化 Π_m．设 $\overline{K}_m$ 是它的佩亚诺核．我们可以证明

$$\int_a^b[\overline{K}_m(t)]^2\mathrm{d}t\leqslant\int_a^b[K_m(t)]^2\mathrm{d}t \tag{5}$$

泛函

$$\theta=\varphi\circ L-\psi=(\varphi-\psi)-(\varphi-\varphi\circ L)$$

的佩亚诺核是 $\overline{\overline{K}}_m=K_m-\overline{K}_m$，并且

$$\overline{\overline{K}}_m(t)=\frac{1}{m!}\theta_x[(x-t)_+^m] \tag{6}$$

如果$\{s_0,s_1,\cdots,s_n\}$是自然样条空间的一组具有基性质的基，那么 $s_i(t_j)=\delta_{ij}$ 并且 L 具有形式

$$Lf=\sum_{i=0}^{n}f(t_i)s_i$$

由此可得 θ 具有形式

$$\theta(f)=\varphi(Lf)-\psi(f)=\sum_{i=0}^{n}f(t_i)\varphi(s_i)-\sum_{i=0}^{n}c_if(t_i)=\sum_{i=0}^{n}\gamma_if(t_i)$$

从而，对于 $\overline{\overline{K}}_m$ 我们有等式

$$\overline{\overline{K}}_m(t)=\frac{1}{m!}\sum_{i=0}^{n}\gamma_i(t_i-t)_+^m \tag{7}$$

选取函数 g 使得 $g^{(m+1)}=\overline{\overline{K}}_m$，则 $g^{(2m+1)}=\overline{\overline{K}}_m^{(m)}$，而且这是一个阶梯函数或者 0 次样条函数．所以，g 自身是具有结点 t_0，t_1，…，t_n 的 $2m+1$ 次样条函数．事实上，g 是一个自然样条．为了证明这一点．我们注意到，根据(7)式，当 $t\geqslant b$ 时有 $\overline{\overline{K}}_m(t)=0$．因此，$g^{(m+1)}(t)=0$，$t\geqslant b$．当 $t\leqslant a\leqslant x$ 时，由(6)式知：

$$\overline{\overline{K}}_m(t)=\frac{1}{m!}\theta_x[(x-t)^m]$$

因为 θ 零化 Π_m，所以当 $-\infty<t\leqslant a$ 时 $\overline{\overline{K}}_m(t)=g^{(m+1)}(t)=0$．因为 g 是自然样条，所以 $Lg=g$．从而

$$\int_a^b\overline{K}_m\overline{\overline{K}}_m\mathrm{d}t=\int_a^b\overline{K}_m g^{(m+1)}\mathrm{d}t=(\varphi-\varphi\circ L)(g)=0$$

517

因为

$$\int_a^b K_m^2\mathrm{d}t=\int_a^b(\overline{K}_m+\overline{\overline{K}}_m)^2\mathrm{d}t=\int_a^b(\overline{K}_m^2+2\overline{K}_m\overline{\overline{K}}_m+\overline{\overline{K}}_m^2)\mathrm{d}t$$

$$=\int_a^b\overline{K}_m^2\mathrm{d}t+\int_a^b\overline{\overline{K}}_m^2\mathrm{d}t\geqslant\int_a^b\overline{K}^2\mathrm{d}t$$

由此可得到(5)式．根据函数的内积，这些计算表明 $\langle\overline{K}_m, \overline{\overline{K}}_m\rangle=0$，由毕达哥拉斯法则，有

$$\|K_m\|_2^2=\|\overline{K}_m+\overline{\overline{K}}_m\|_2^2=\|\overline{K}_m\|_2^2+\|\overline{\overline{K}}_m\|_2^2\geqslant\|\overline{K}_m\|_2^2$$ ■

例 3 设

$$\varphi(f)=\int_{-1}^{1}f(x)\mathrm{d}x$$

$$\psi(f)=c_1f(-1)+c_2f(0)+c_3f(1)$$

只考虑在 Π_1 上与 φ 一致的泛函 ψ，求出 Sard 意义下 φ 的最佳逼近．

解 我们要求出这种情形下的自然样条插值算子 L．可以证明

$$(Lf)(x)=a_0+a_1x+b_0(x+1)_+^3+b_1(x)_+^3+b_2(x-1)_+^3$$

其中

$$a_0=[-f(-1)+6f(0)-f(1)]/4$$

$$a_1=[-5f(-1)+6f(0)-f(1)]/4$$

$$b_0=b_2=-b_1/2=[f(-1)-2f(0)+f(1)]/4$$

因而最佳公式是

$$\psi(f)=\varphi(Lf)=\int_{-1}^{1}(Lf)(x)\mathrm{d}x$$

$$=2a_0+4b_0+\frac{1}{4}b_1$$

$$=\frac{3}{8}f(-1)+\frac{5}{4}f(0)+\frac{3}{8}f(1)$$ ■

习题 7.6

1. 寻找一个逼近公式

518

$$f'(x) \approx c_0 f(-1) + c_1 f(0) + c_2 f(1)$$

其中 x 是 $(-1, 1)$ 中给定的点. 找出为使公式在 Π_1 上成立，系数所满足的充分必要条件. 求出 Sard 意义下的最佳逼近. 当达到极小值时计算 $\int_{-1}^{1} K_1(t)^2 \mathrm{d}t$.

2. 对于例 3，求出误差公式

$$v \mid \varphi(f) - \psi(f) \mid \leqslant c \| f'' \|_2$$

中的最佳常数.

3. 求出插值函数 f $(f(t_i) = \lambda_i$，$0 \leqslant i \leqslant n)$，它使得积分 $\int_{t_0}^{t_n} (f'')^2 \mathrm{d}t$ 极小. 假设 f'' 是分段连续的.

4. 应用佩亚诺核定理于泛函

$$\varphi(f) = f(x+h) - f(x) - (h/2)[3f'(x) - f'(x-h)]$$

$(x, h$ 是固定的$)$，证明 $\varphi(f) = (5/12)h^3 f'''(\xi)$，其中 $x-h < \xi < x+h$.

5. 利用佩亚诺核定理得到辛普森法则的结果：

$$\int_0^2 f(x)\mathrm{d}x = \frac{1}{3}[f(0) + 4f(1) + f(2)] - \frac{1}{90}f^{(4)}(\xi)$$

6. 利用 Schoenberg 定理证明

$$\left| f(2) + \frac{1}{4}f(0) - \frac{7}{8}f(1) - \frac{3}{8}f(3) \right| \leqslant \sqrt{\frac{5}{48}} \| f'' \|_2$$

7.7 伯努利多项式和欧拉-麦克劳林公式

龙贝格求积算法的正确性取决于欧拉-麦克劳林公式，即 7.4 节中的(6)式. 我们先给出伯努利多项式的一些性质，然后推导这个公式.

伯努利多项式

伯努利多项式形成了一个无穷序列，它具有许多有用的性质. 最初的几个伯努利多项式是

$$B_0(t) = 1$$

$$B_1(t) = t - \frac{1}{2}$$

$$B_2(t) = t^2 - t + \frac{1}{6}$$

$$B_3(t) = t^3 - \frac{3}{2}t^2 + \frac{1}{2}t$$

它们是由下列等式定义的

$$\sum_{k=0}^{n} \binom{n+1}{k} B_k(t) = (n+1)t^n \tag{1}$$

回忆

519

$$\binom{n}{m} = \frac{n!}{m!(n-m)!}$$

例如，当 $n=0$，等式(1)表明 $B_0(t)=1$. 当 $n=1$ 时，等式(1)表明 $B_0(t)+2B_1(t)=2t$，由此可得 $B_1(t)=t-(1/2)$. 由于(1)式中 B_n 的系数是 $n+1$，所以我们总可以根据 B_0，B_1，…，B_{n-1} 求出 B_n. 下面的定理列举了我们所需要的伯努利多项式的性质.

定理 1(伯努利多项式性质定理) 伯努利多项式具有下列性质：

1. $B_n'=nB_{n-1}(n\geqslant 1)$.
2. $B_n(t+1)-B_n(t)=nt^{n-1}(n\geqslant 2)$.
3. $B_n(t)=\sum_{k=0}^{n}\binom{n}{k}B_k(0)t^{n-k}$.
4. $B_n(1-t)=(-1)^nB_n(t)$.

证明 用数学归纳法证明性质 1. 根据事实 $B_1(t)=t-(1/2)$ 和 $B_0(t)=1$ 知，$n=1$ 时结论成立. 现在假设 $B'_k=kB_{k-1}$，$k=1, 2, \cdots, n-1$. 那么根据(1)式及其微分，我们得到

$$\sum_{k=1}^{n}\binom{n+1}{k}B'_k(t)=n(n+1)t^{n-1}=(n+1)\sum_{k=0}^{n-1}\binom{n}{k}B_k(t) \tag{2}$$

根据归纳假设，(2)式可写为

$$(n+1)B_n'+\sum_{k=1}^{n-1}\binom{n+1}{k}kB_{k-1}=(n+1)\sum_{k=0}^{n-1}\binom{n}{k}B_k$$

等式两边同除以 $n+1$ 并且利用恒等式

$$\frac{k}{n+1}\binom{n+1}{k}=\binom{n}{k-1}$$

得到

$$B'_n+\sum_{k=1}^{n-1}\binom{n}{k-1}B_{k-1}=\sum_{k=0}^{n-1}\binom{n}{k}B_k$$

消去相同的项便得到所要证明的性质 1. 为了证明性质 2. 重复利用性质 1. 得到

$$\begin{aligned}B''_n(t)&=n(n-1)B_{n-2}(t)\\B'''_n(t)&=n(n-1)(n-2)B_{n-3}(t)\\&\vdots\\B_n^{(k)}(t)&=n(n-1)\cdots(n-k+1)B_{n-k}(t)\end{aligned}$$

从而，B_n 的泰勒级数是

520

$$B_n(t+h)=\sum_{k=0}^{n}\frac{1}{k!}B_n^{(k)}(t)h^k=\sum_{k=0}^{n}\binom{n}{k}B_{n-k}(t)h^k \tag{3}$$

现在取 $h=1$. 利用恒等式

$$\binom{n}{k}=\binom{n}{n-k}$$

以及(1)式可得到性质 2：

$$B_n(t+1)=\sum_{k=0}^{n}\binom{n}{k}B_k(t)=B_n(t)+\sum_{k=0}^{n-1}\binom{n}{k}B_k(t)$$

$$= B_n(t) + nt^{n-1}$$

在(3)式中取 $t=0$ 及 $h=t$，我们得到性质 3.

最后证明性质 4. 在性质 2 中用 $-t$ 代替 t，得到

$$B_n(1-t) - B_n(-t) = n(-1)^{n-1}t^{n-1}$$
$$= (-1)^{n-1}[B_n(t+1) - B_n(t)]$$

把它写为下列形式

$$(-1)^n B_n(t+1) - B_n(-t) = (-1)^n B_n(t) - B_n(1-t) \equiv F(t)$$

我们把它看作形如 $F(t+1)=F(t)$ 的等式，这说明 F 是周期为 1 的函数. 因为 F 是一个多项式，所以它一定是常数. 因此，

$$(-1)^n B_n(t) - B_n(1-t) = c_n \tag{4}$$

再利用性质 1 和(4)式的微分，得

$$(-1)^n B'_n(t) + B'_n(1-t) = (-1)^n nB_{n-1}(t) + nB_{n-1}(1-t) = 0$$

这与性质 4 等价. ■

引理 1(伯努利多项式引理) 函数 $G(t)=B_{2n}(t)-B_{2n}(0)$ 在开区间(0, 1)中没有零点.

证明 在定理 1 的性质 2 和性质 4 中，令 $t=0$ 得到:

$$B_n(0) = B_n(1) = (-1)^n B_n(0)$$

从而，$B_3(0)=B_5(0)=B_7(0)=\cdots=0$. 现在假设 G 在区间(0, 1)中有一个零点. 由于 $G(0)=G(1)=0$，根据罗尔定理可知 G' 在(0, 1)中有两个零点. 因为 B_{2n-1} 是 G' 的倍数，所以 B_{2n-1} 在(0, 1)中有两个零点. 但 B_{2n-1} 在点 0 和 1 取零值，所以 B'_{2n-1} 在(0, 1)中有 3 个零点. 因而，B_{2n-2} 在(0, 1)中有 3 个零点. 这样继续下去，我们得知对于奇数指标 $k<2n$，B_k 在(0, 1)中至少有两个零点. 由此，B_3 除了两个零点 0 和 1 之外，它在(0, 1)中还有两个零点. 因为 B_3 是一个三次多项式，这显然是不可能的. ■

521

定理 2(欧拉-麦克劳林公式) 若函数 f 在[0, 1]中有 $2n$ 阶连续导数，则

$$\int_0^1 f(t)\mathrm{d}t = \frac{1}{2}[f(0)+f(1)] - \sum_{k=0}^{n-1}\frac{b_{2k}}{(2k)!}[f^{(2k-1)}(1) - f^{(2k-1)}(0)] + R$$

其中

$$\begin{cases} b_k = B_k(0) \\ R = -\dfrac{b_{2n}}{(2n)!}f^{(2n)}(\xi) \quad (0<\xi<1) \end{cases}$$

证明 根据定理 1 中性质 1，我们有

$$B_n(t) = \frac{1}{n+1}B'_{n+1}(t)$$

利用该公式以及分部积分法，我们给出

$$\int_0^1 f(t)\mathrm{d}t = \int_0^1 f(t)B_0(t)\mathrm{d}t = B_1(t)f(t)\Big|_0^1 - \int_0^1 B_1(t)f'(t)\mathrm{d}t \tag{5}$$

因为 $B_1(1)=1/2$ 及 $B_1(0)=-1/2$，(5)式可写为

$$\int_0^1 f(t)\mathrm{d}t = \frac{1}{2}[f(1)+f(0)] - \int_0^1 B_1(t)f'(t)\mathrm{d}t$$

此式中的积分也可由分部积分法得出，其结果是

$$\int_0^1 f(t)\mathrm{d}t = \frac{1}{2}[f(1)+f(0)] - \frac{b_2}{2}[f'(1)-f'(0)] + \frac{1}{2}\int_0^1 B_2(t)f''(t)\mathrm{d}t$$

可以继续这个过程. 根据前面给出的公式

$$B_n(0) = B_n(1) = b_n$$
$$b_3 = b_5 = b_7 = \cdots = 0$$

$2n$ 步以后，我们有

$$\int_0^1 f(t)\mathrm{d}t = \frac{1}{2}[f(1)+f(0)] - \sum_{k=1}^{n}\frac{b_{2k}}{(2k)!}[f^{(2k-1)}(1)-f^{(2k-1)}(0)] + \frac{1}{(2n)!}\int_0^1 B_{2n}(t)f^{(2n)}(t)\mathrm{d}t$$

和式中的最后一项可以表示为

$$\frac{b_{2n}}{(2n)!}[f^{(2n-1)}(1)-f^{(2n-1)}(0)] = \frac{b_{2n}}{(2n)!}\int_0^1 f^{(2n)}(t)\mathrm{d}t$$

522

现在除了余项是

$$R = \frac{1}{(2n)!}\int_0^1 [B_{2n}(t)-b_{2n}]f^{(2n)}(t)\mathrm{d}t \qquad (6)$$

之外，可以把公式写成定理中叙述的形式. 根据伯努利多项式引理 1，函数 $B_{2n}(t)-b_{2n}$ 在[0，1]中不改变符号. 因此，将积分中值定理应用于(6)式，R 可以写为

$$R = \frac{1}{(2n)!}f^{(2n)}(\xi)\int_0^1 [B_{2n}(t)-b_{2n}]\mathrm{d}t$$

因为 $B_{2n}=B'_{2n+1}/(2n+1)$ 并且在 $t=0$ 和 $t=1$ 时 $B_{2n+1}(t)=0$，所以最后我们得到

$$R = -\frac{b_{2n}}{(2n)!}f^{(2n)}(\xi)$$

■

习题 7.7

1. 证明：$B_n(t)$ 的首项是 t^n 而 t^{n-1} 项的系数是 $-n/2$.
2. 利用定理 1 中的性质 2，证明：

$$1^p + 2^p + \cdots + n^p = [B_{p+1}(n+1) - B_{p+1}(0)]/(p+1)$$

3. 证明：当 n 是奇数时，函数 $B_n(x)-B_n(0)$ 在 1/2 处有单个零点.
4. 证明：数 $B_0(0)$、$B_2(0)$、$B_4(0)$、…的符号交替变化.
5. 证明：对于偶数 n，B_n 在(0，1)中至少有两个根；对于奇数 n，B_n 在(0，1)中至少有一个根.
6. 证明：二项式系数 $\binom{n}{m}$ 是整数. 提示：可以首先证明帕斯卡定律：

$$\binom{n}{m-1} + \binom{n}{m} = \binom{n+1}{m}$$

523

第8章　常微分方程数值解

8.0　概述

本章所关心的是包含常微分方程的数值解问题．中心问题是当一个点在解曲线上是已知时求解单个一阶方程．后面几节讨论方程组、高阶方程和两点边值问题．

8.1　解的存在性和唯一性

我们的模型是下列初值问题

$$\begin{cases} x' = f(t,x) \\ x(t_0) = x_0 \end{cases} \tag{1}$$

这里 x 是 t 的未知函数，我们希望从(1)中式给出的信息去构造它，其中 $x'=\mathrm{d}x(t)/\mathrm{d}t$．(1)式中第2个等式指定函数 $x(t)$ 的一个特定值．第1个等式给出曲线 x 在任意点 t 上的斜率．当然，函数 f 必须指定．作为具体的例子，可取　524

$$\begin{cases} x' = x\tan(t+3) \\ x(-3) = 1 \end{cases} \tag{2}$$

我们要在一个包含初始点 t_0 的区间上确定 $x(t)$．容易验证这个初值问题的解析解是 $x(t)=\sec(t+3)$．因为 $\sec t$ 在 $t=\pm\pi/2$ 时变成无穷，所以我们的解仅对 $-\pi/2<t+3<\pi/2$ 成立．(2)式中的例子是例外的，因为它有一个简单的解析解，从中能容易地计算解的数值．对(1)中类型的问题，典型情况是不能利用解析解，必须使用数值方法．

8.1.1　存在性

形如(1)式的每个初值问题是否有解？回答是否定的．对 f 必须作某些假设，尽管这样我们可期待仅仅在 $t=t_0$ 的邻域中存在解．作为可能发生的例子，考虑

$$\begin{cases} x' = 1+x^2 \\ x(0) = 0 \end{cases} \tag{3}$$

解曲线从 $t=0$ 出发，斜率为1；即 $x'(0)=1$．因为斜率是正的，$x(t)$ 靠近 $t=0$ 时递增，所以表达式 $1+x^2$ 也递增，因此 x' 递增．因为 x 和 x' 都递增且有关系 $x'=1+x^2$，所以我们可期待一个有垂直渐近线的解．事实上，因为(3)的解析解是 $x(t)=\tan t$，所以这种情况在 $t=\pi/2$ 处出现．

定理1(第一存在性定理，初值问题)　*若 f 在中心为 $(t_0,\ x_0)$ 的矩形*

$$R = \{(t,x):|t-t_0|\leqslant\alpha,\quad |x-x_0|\leqslant\beta\} \tag{4}$$

内连续，则对于 $|t-t_0|\leqslant\min(\alpha,\ \beta/M)$ 初值问题(1)有解 $x(t)$，其中 M 是 $|f(t,\ x)|$ 在矩形 R 内的最大值．

例1　证明初值问题

$$\begin{cases} x' = (t+\sin x)^2 \\ x(0) = 3 \end{cases}$$

525 在区间$-1\leqslant t\leqslant 1$上有解.

解 在此例中，$f(t, x)=(t+\sin x)^2$，$(t_0, x_0)=(0, 3)$. 在矩形

$$R=\{(t,x):|t|\leqslant\alpha,\ |x-3|\leqslant\beta\}$$

内f的值以$|f(t, x)|\leqslant(\alpha+1)^2\equiv M$为界. 我们要$\min(\alpha, \beta/M)\geqslant 1$，故可设$\alpha=1$. 则$M=4$，并且我们的目标通过取$\beta\geqslant 4$来实现. 存在性定理表明在区间$|t|\leqslant\min(\alpha, \beta/M)=1$上存在初值问题的解. ■

8.1.2 唯一性

即使f连续，也有可能出现初值问题(1)没有唯一解的情况. 这种现象的一个简单例子由下列问题给出

$$\begin{cases}x'=x^{2/3}\\x(0)=0\end{cases}$$

显然，0函数，即$x(t)\equiv 0$是这个问题的解. 另一个解是函数$x(t)=\frac{1}{27}t^3$.

为证明初值问题(1)在$t=t_0$的邻域中有唯一解，对f稍为多作一点假设是必要的. 关于这个问题这里有一个有用的定理.

定理2(唯一性定理，初值问题) *若f和$\partial f/\partial x$在(4)式定义的矩形R内连续，则初值问题(1)在区间$|t-t_0|<\min(\alpha, \beta/M)$内有唯一解.*

在定理1和定理2这两个定理中，在t轴上表明解存在的区间可能小于我们定义$f(t, x)$的矩形底部. 下面的定理具有不同的类型，它允许我们在指定的区间$[a, b]$上推断解的存在性和唯一性.（见Henrici[1962，第15页].）

定理3(存在性定理，初值问题) *若f在带$a\leqslant t\leqslant b$，$-\infty<x<+\infty$内连续并满足不等式*

$$|f(t,x_1)-f(t,x_2)|\leqslant L|x_1-x_2| \tag{5}$$

则初值问题(1)在区间$[a, b]$上有唯一解.

不等式(5)称为关于第2个变量的**利普希茨条件**. 对于单变量函数，这种条件简化为

$$|g(x_1)-g(x_2)|\leqslant L|x_1-x_2| \tag{6}$$

526 可以直接看出这个条件比连续性更强，因为若x_2逼近x_1时，(6)式的右边接近于0，并且这迫使$g(x_2)$逼近$g(x_1)$. 条件(6)比具有有界导数弱些. 当然，若$g'(x)$处处存在且按模不超过L，则由中值定理，得

$$|g(x_1)-g(x_2)=|g'(\xi)||x_1-x_2|\leqslant L|x_1-x_2|$$

例2 说明函数$g(x)=\sum_{i=1}^{n}a_i|x-w_i|$满足具有常数$L=\sum_{i=1}^{n}|a_i|$的利普希茨条件.

解

$$\begin{aligned}|g(x_1)-g(x_2)|&=\left|\sum_{i=1}^{n}a_i|x_1-w_i|-\sum_{i=1}^{n}a_i|x_2-w_i|\right|\\&=\left|\sum_{i=1}^{n}a_i\{|x_1-w_i|-|x_2-w_i|\}\right|\\&\leqslant\sum_{i=1}^{n}|a_i|\,||x_1-w_i|-|x_2-w_i||\\&\leqslant\sum_{i=1}^{n}|a_i||x_1-x_2|=L|x_1-x_2|\end{aligned}$$

■

习题 8.1

1. 求初值问题

$$\begin{cases} x' = x^{1/3} \\ x(0) = 0 \end{cases}$$

的两个解. 提示：试验 $x=ct^k$ 或观察方程是可分离的.

2. a. 利用初值问题存在性定理 1 来预测在怎样的区间中初值问题(3)的解存在. 求出最大的区间.

   b. 对初值问题(2)重复 a.

3. 说明 $x=-t^2/4$ 和 $x=1-t$ 是初值问题

$$\begin{cases} 2x' = \sqrt{t^2+4x}-t \\ x(2) = -1 \end{cases}$$

的解. 为什么这与初值问题唯一性定理 2 不矛盾?

4. 就下列特殊情况求解初值问题 $x'=f(t, x)$，$x(0)=0$.

   a. $f(t, x)=t^3$

   b. $f(t, x)=(1-t^2)^{-1/2}$

   c. $f(t, x)=(1+t^2)^{-1}$

   d. $f(t, x)=(t+1)^{-1}$

527

5. 就下列情况求解初值问题 $x'=f(t, x)$，$x(0)=0$. 当 $\mathrm{d}x/\mathrm{d}t\neq 0$ 时，利用 $\mathrm{d}t/\mathrm{d}x=(\mathrm{d}x/\mathrm{d}t)^{-1}$.

   a. $f(t, x)=x^{-2}$

   b. $f(t, x)=1+x^2$

   c. $f(t, x)=(\sin x+\cos x)^{-1}$

6. 利用初值问题存在性定理 1，证明初值问题

$$\begin{cases} x' = \sqrt{|x|} \\ x(0) = 0 \end{cases}$$

在整条实线上有解.

7. 利用初值问题存在性定理 1 证明初值问题

$$\begin{cases} x' = \tan x \\ x(0) = 0 \end{cases}$$

在区间 $|t|<\pi/4$ 内有解.

8. 设 f 是定义在整个$\mathbb{R}$上的单变量连续函数，$M(r)$表示 $|x|\leqslant r$ 时 $|f(x)|$ 的最大值. 若 $r\to\infty$时，$M(r)=o(r)$，则初值问题

$$\begin{cases} x' = f(x) \\ x(0) = 0 \end{cases}$$

在整个$\mathbb{R}$上有解. 证明这个命题.

9. 证明初值问题

$$\begin{cases} x' = t^2+\mathrm{e}^x \\ x(0) = 0 \end{cases}$$

在区间 $|t|\leqslant 0.351$ 内有唯一解.

10. 证明：若 $f(t, x)$连续且在域 $a\leqslant t\leqslant b$，$-\infty<x<+\infty$内有界，则初值问题

$$\begin{cases} x' = f(t,x) \\ x(a) = \alpha \end{cases}$$

在区间 $a\leqslant t\leqslant b$ 内有解.

11. 设 R 表示 tx 平面内由 $|t-t_0|\leqslant\alpha$, $|x-x_0|\leqslant\beta$ 定义的矩形，f 是定义在这个矩形上并满足 $\beta\geqslant\alpha|f(t,x)|$ 的连续函数. 证明初值问题 $x'=f(t,x)$, $x(t_0)=x_0$ 在区间 $|t-t_0|\leqslant\alpha$ 上有解.

12. 证明初值问题

$$\begin{cases}x'=1+x+x^2\cos t\\x(0)=0\end{cases}$$

528 在区间 $-1/3\leqslant t\leqslant 1/3$ 内有解.

13. 说明初值问题

$$\begin{cases}x'=\sqrt{|x|}\\x(0)=0\end{cases}$$

有两个解，并指出为什么不应用初值问题唯一性定理2.

14. 证明初值问题

$$\begin{cases}x'=tx^{2/3}\\x(0)=1\end{cases}$$

在区间 $-2\leqslant t\leqslant 2$ 内有解. 是否存在多个解?

15. 证明初值问题

$$\begin{cases}x'=2te^{-x}\\x(0)=0\end{cases}$$

在区间 $-\infty<t<+\infty$ 内有唯一解. 试问，在允许引出这个结论的存在性定理中是否存在 α 和 β 的一种选择?

16. 设 $f(t,x)$ 在 $|t|\leqslant 3$, $|x|\leqslant 4$ 定义的矩形上连续，并假设在这个矩形内一切点上 $|f(t,x)|\leqslant 7$. 试问初值问题

$$\begin{cases}x'=f(t,x)\\x(0)=0\end{cases}$$

必定有解的最大区间是什么?

17. 求出能保证初值问题

$$\begin{cases}x'=\sec x\\x(0)=0\end{cases}$$

有唯一解的区间.

18. 利用本节的3个定理中的每个定理预测下列问题

$$\begin{cases}x'=x^2\\x(0)=1\end{cases}$$

有解的区域，然后明确地求解该问题，把理论结果和实际计算作比较.

19. 证明初值问题

$$\begin{cases}x'=1+x^2\\x(0)=0\end{cases}$$

在区间[−1, 1]内有解. 证明这个例子不满足定理3的假设. 说明为什么这个例子与定理3不矛盾.

计算机习题 8.1

529 利用符号操作程序，求微分方程 $x'+x=(1+e^t)^{-1}$ 的解.

8.2 泰勒级数方法

在微分方程的数值解中，我们很少期望直接得到公式解，给出 $x(t)$ 为 t 的函数. 代之，通

常构造下列形式的函数值表

t_0	t_1	t_2	t_3	$\cdots$	t_m
x_0	x_1	x_2	x_3	$\cdots$	x_m

(1)

这里，x_i 是在 t_i 的精确解 $x(t_i)$ 的近似计算值，根据如同(1)这样的表，可构造样条函数或其他的近似函数. 然而，求解常微分方程的大多数数值方法首先产生这样的一个表.

我们再次考察初值问题

$$\begin{cases} x' = f(t,x) \\ x(t_0) = x_0 \end{cases} \tag{2}$$

其中，f 是一个给定的两个变量的函数，$(t_0,\ x_0)$ 是一个单独的解曲线经过的已知点. (2)式的解是一个函数 $x \mapsto x(t)$，使得对 t_0 的某个邻域内的一切 t 有 $\mathrm{d}x(t)/\mathrm{d}t = f(t,\ x(t))$ 且 $x(t_0) = x_0$.

8.2.1 实例

泰勒级数方法必须假定 f 的各种偏导数存在. 为说明此方法，我们举个实例：

$$\begin{cases} x' = \cos t - \sin x + t^2 \\ x(-1) = 3 \end{cases} \tag{3}$$

方法的要点是关于 x 的泰勒级数，我们记为

$$x(t+h) = x(t) + hx'(t) + \frac{h^2}{2!}x''(t) + \frac{h^3}{3!}x'''(t) + \frac{h^4}{4!}x^{(4)}(t) + \cdots \tag{4}$$

这里出现的导数可从(3)式中的微分方程得到. 它们是

$$x'' = -\sin t - x'\cos x + 2t$$
$$x''' = -\cos t - x''\cos x + (x')^2\sin x + 2$$
$$x^{(4)} = \sin t - x'''\cos x + 3x'x''\sin x + (x')^3\cos x$$

此刻，我们的耐性逐渐消失，决定仅使用公式(4)中按升次直到包含 h^4 的那些项. 那些不被包含的从 h^5 开始的项共同构成方法中固有的**截断误差**. 所得的数值方法称为四阶方法. (如果使用按升次直到包含 $h^n x^{(n)}(t)/n!$ 的那些项，那么泰勒级数方法的阶为 n.) 注意，在求诸如 $\sin x$ 的项关于 t 的微分中，我们必须把它看作 $\mathrm{d}\{\sin[x(t)]\}/\mathrm{d}t$ 并利用微分的链式法则. 当然，这说明关于 x''，x'''，…公式的复杂性. 我们可以执行各种代换来得到右边不含 x 的导数 x''，x'''，…的公式. 如果公式是按列出的次序使用的话就不必做这项工作. 事实上，它们是递归的. 530

下面是求解初值问题(3)的一个算法，从 $t=-1$ 开始，步长为 $h=0.01$. 我们要求 t 区间 $[-1,\ 1]$ 上的一个解，因而要执行 200 步.

```
input M←200; h←0.01; t← −1.0; x←3.0
output 0, t, x
for k = 1 to M do
    x′←cost − sinx + t²
    x″← −sint − x′cosx + 2t
    x‴← −cost − x″cosx + (x′)²sinx + 2
    x⁽⁴⁾←sint + ((x′)³ − x‴)cosx + 3x′x″sinx
```

```
        x←x+h(x'+h/2(x''+h/3(x'''+h/4(x^(4)))))
        t←t+h
        output k, t, x
    end do
```

关于计算解中的误差能够说些什么呢? 在每一步，因为不包含泰勒级数中涉及 h^5，h^6，… 的项，所以**局部截断误差**是 $\mathcal{O}(h^5)$. 因此，当 $h\to 0$ 时，局部误差的情况类似于 Ch^5. 遗憾的是，我们不知道 C. 因为 $h=10^{-2}$，所以 h^5 是 10^{-10}. 因而侥幸地，每一步中的误差粗略地具有 10^{-10} 的量级. 几百步后，这些小的误差可能累加起来并且损坏数值解. 在每一步(除第 1 步外)，$x(t_k)$ 的估计 x_k 已经包含误差，进一步的计算继续增加这些误差. 因此，要小心在微分方程的数值解中盲目地采用所有的十进制数字.

当前面的算法被编程并运行时，在 $t=1$ 上解是 $x_{200}=6.421\,94$. 下面是来自计算机程序输出的一个样本:

k	t	x
0	−1.000 00	3.000 00
1	−0.990 00	3.014 00
2	−0.980 00	3.028 03
3	−0.970 00	3.042 09
4	−0.960 00	3.056 17
5	−0.950 00	3.070 28
6	−.940 00	3.084 43
7	−.930 00	3.098 61
⋮	⋮	⋮
196	0.960 00	6.365 66
197	0.970 00	6.379 77
198	0.980 00	6.393 86
199	0.990 00	6.407 91
200	1.000 00	6.421 94

531

在随后的计算机运行中，微分方程用 x_{200} 这个值作为初始条件且用 $h=-0.01$ 来积分. 第 2 次计算机运行的结果是在 $t=-0.999\,99$ 上 $x_{200}\approx 3.000\,00$. 它与原来的初始值几乎相同，这使我们认为数值解所示的数的 6 位有效数字是准确的.

在刚才讨论的例子中，估计数值解每步中的局部截断误差是不难的. 为此，我们回顾泰勒级数(4)中的误差项具有形式

$$E_n=\frac{1}{(n+1)!}h^{n+1}x^{(n+1)}(t+\theta h) \qquad (0<\theta<1)$$

这是当包含在和式中的 h 最后的幂是 h^n 时出现的误差. 在此例中，取 $n=4$，$h=0.01$. 我们可用简单的有限差分逼近估计 $x^{(5)}(t+\theta h)$ 并得出结论

$$E_4 \approx \frac{1}{5!}h^5\left[\frac{x^{(4)}(t+h)-x^{(4)}(t)}{h}\right] = \frac{h^4}{120}\left[x^{(4)}(t+h)-x^{(4)}(t)\right] \tag{5}$$

用一点点额外的编程，E_4 的这个估计可合并到算法中并算出．这里我们不给出细节，而是它作为计算机习题 8.2.19．当这部分加入程序时，计算的输出表明估计 E_4 的绝对值决不会超过 3.42×10^{-11}．程序在类似于 Marc-32 那样的计算机上以双精度运行．

8.2.2 权衡利弊

泰勒级数方法的优缺点是什么？其缺点是这个方法依赖于给定的微分方程的反复求导(除非我们只打算用一阶方法)．因此，在解曲线经过的 tx 平面的区域内函数 $f(t, x)$ 必须具有偏导数．当然，这样的一个假设对于解的存在性是不必要的．执行初步的分析工作的必要性也是一个突出的缺点．例如，在这个步骤中造成的误差可能被忽略且始终不被发现．最后，各种求导必须被分别地编程．

其优点是方法概念的简单性并且具有非常高精度的潜力．因此，如果能容易地得到 $x(t)$ 的 20 阶导数，则没有什么能阻止我们使用 20 阶的方法(即，按升次直到包含 h^{20} 的那些项)．具有这样高的阶，同样的精确度可用较大的步长得到，譬如说，$h=0.2$．穿过给定的区间需要较少的步数，这就减少了计算工作量．另一方面，在每一步中的计算负担会较重．在习题中有一些可以使用高阶泰勒级数方法的例子． 532

近年来，已经可以利用符号操作程序执行非数值类型的各种数学计算的程序．因而可把相当复杂的表达式的微分和积分转换到计算机上进行！当然，这些程序仅仅应用于一个有限制的函数类，但是这类函数是足够广泛的，它包括典型的微积分教材中会遇到的所有函数．使用这种程序，可以毫无困难地使用 20 阶的泰勒级数方法．

8.2.3 误差

在数值求解微分方程中，会产生若干种类型的误差．这些误差可以方便地分类如下：

1. 局部截断误差．
2. 局部舍入误差．
3. 整体截断误差．
4. 整体舍入误差．
5. 总误差．

局部截断误差是当用一个有限过程代替无限过程时，在一步中产生的误差．在泰勒级数方法中，我们用部分和代替 $x(t+h)$ 的无穷泰勒级数．局部截断误差是我们可能选择的任何算法中固有的，它与舍入误差完全无关．当然，**舍入误差**是由计算机的有限精度引起的，它的值与计算机的字长有关(或者与浮点机器数尾数中的位数有关)．

在泰勒级数方法中，如果保留级数中按升次直到包含 h^n 的那些项，则局部截断误差是我们不包含的所有其余项之和．由泰勒定理，对某个 t 附近的点 ξ，这些项可压缩成下列形式的单独的一项

$$\frac{h^{n+1}}{(n+1)!}x^{(n+1)}(\xi)$$

我们说局部截断误差是$\mathcal{O}(h^{n+1})$．这类误差出现在数值解的每一步中．许多局部截断误差全体的累积引起**整体截断**误差．即使所有的计算都用精确的运算，这个误差还是会出现．它与方法有关而与执行计算的计算机无关．若局部截断误差是$\mathcal{O}(h^{n+1})$，则整体截断误差必定是$\mathcal{O}(h^n)$，因为从t_0开始到达任意点T必要的步数是$(T-t_0)/h$.

整体舍入误差是前面的步骤中局部舍入误差的累积．**总误差**是整体截断误差和整体舍入误差之和．若整体截断误差是$\mathcal{O}(h^n)$，则我们说数值方法具有**n阶**.

533

8.2.4 欧拉方法

$n=1$时的泰勒级数方法称为**欧拉方法**．它具有形式

$$x(t+h)=x(t)+hf(t,x)$$

这个公式具有明显的优点，它不需要对f求任何导数．为了得到合意的精度必须采取小的h值从而抵消了这个优点．因为存在性定理能以它为基础，所以该方法作为一个有用的例子仍然具有重大的理论上的重要性．这样的存在性定理见Henrici[1962，第15-25页].

8.2.5 延迟微分方程

在一些实际问题中出现了一种特殊类型的微分方程，称为**延迟微分方程**或**具有延迟变量的微分方程**．人口模型以及混合问题通常有此特征，就是说$x'(t)$的值与x在t的前面值上的函数值有关．例如，我们可能有

$$x'(t)=f(x(t-1))$$

若知道x在$t-1$上的值，微分方程就能够计算$x'(t)$的值．为了从$t=0$开始积分微分方程，我们需要在$t=-1$开始的$x(t)$的变化情况．因此，必须提供$x(t)$在区间$[-1,0]$上的值作为初值.

这一类型的一个特殊的和明确的问题可能是

$$\begin{cases}x'(t)=x(t-1) & (t\geqslant 0)\\ x(t)=t^2 & (-1\leqslant t\leqslant 0)\end{cases}\tag{6}$$

第2个等式给出所需要的$x(t)$的初值．若t限定于区间$[0,1]$中，则$t-1$在$[-1,0]$中，因此

$$\begin{cases}x'(t)=x(t-1)=(t-1)^2 & (0\leqslant t\leqslant 1)\\ x(0)=0\end{cases}$$

通过积分并提供一个适当的积分常数，容易得到这个解：

$$x(t)=\frac{1}{3}(t-1)^3+\frac{1}{3}\qquad(0\leqslant t\leqslant 1)$$

如果解被延拓到下一个区间$[1,2]$中，则可以采取类似于第一步的另一步．因此，对于$[1,2]$中的t，我们有

$$\begin{cases}x'(t)=x(t-1)=\dfrac{1}{3}(t-2)^3+\dfrac{1}{3} & (1\leqslant t\leqslant 2)\\ x(1)=\dfrac{1}{3}\end{cases}$$

534

这个方程的解是

$$x(t)=\frac{1}{12}(t-2)^4+\frac{1}{3}t-\frac{1}{12} \qquad (1\leqslant t\leqslant 2)$$

通过类似的计算，解可以无限地朝右边延拓.

对更复杂的方程，例如

$$x'(t)=\sin[x(t-1)^3]+\log[x(t)+t^5]$$

我们必须求助于数值方法. 泰勒级数方法是可利用的，但不是没有某些缺点，譬如确定导数，这些缺点可能是复杂的而且有误差倾向. 例如，考虑

$$\begin{cases} x'(t)=2x(t-1)+x(t) & (t>0) \\ x(t)=t^2 & (-1\leqslant t\leqslant 0) \end{cases} \tag{7}$$

为了在区间[0, 1]中求解，我们利用短的泰勒展开式

$$x(t+h)=x(t)+hx'(t)+\frac{h^2}{2}x''(t)+\frac{h^3}{6}x'''(t)$$

以长度为 h 的步长向前进行求解. 在泰勒级数方法中，照例必须利用微分方程提供 x'，x'' 和 x''' 的公式. 在此例中，如果只考虑区间[0, 1]，则我们可利用

$$x'(t)=2x(t-1)+x(t)=2(t-1)^2+x(t)$$
$$x''(t)=2x'(t-1)+x'(t)=4(t-2)^2+2(t-1)^2+x'(t)$$
$$x'''(t)=2x''(t-1)+x''(t)=8(t-3)^2+8(t-2)^2+2(t-1)^2+x''(t)$$

数值解产生区间[0, 1]中的离散点上 $x(t)$ 的值. 同时，为了在下一个区间内使用，我们必须存放这些相同的离散点上的 $x'(t)$，$x''(t)$ 和 $x'''(t)$ 的值. 若不改变 h 的值，则利用适当的存储值我们能以同样的方法在每个区间上向前进行求解.

关于延迟微分方程的理论可以参见 Driver[1977]、Kuang[1993]以及 Diekmann, Van Gils, Verduyn Lunel, and Walther[1995]的著作. 另外，Willé and Baker[1992]给出了一个用于求解延迟微分方程系统的称为 DELSOL 的代码.

习题 8.2

1. 在计算机习题 8.2.3 中，必须计算函数 e^{-t^2} 的导数. 证明这个函数的 n 阶导数具有形式 $e^{-t^2}P_n(t)$，其中多项式 P_n 是由下列公式递归地决定的.

$$\begin{cases} P_0=1 \\ P_{n+1}(t)=P'_n(t)-2tP_n(t) \end{cases}$$

535

例如，证明

$$P_4(t)=12-48t^2+16t^4$$
$$P_5(t)=-120t+160t^3-32t^5$$

2. 验证初值问题

$$\begin{cases} x'=\sqrt{x} \\ x(0)=0 \end{cases}$$

的解是 $x(t)=t^2/4$. 应用一阶泰勒级数方法并说明为什么数值解与解 $t^2/4$ 不同.

3. 用二阶泰勒级数方法一步解微分方程

$$\begin{cases} x'=-tx^2 \\ x(0)=2 \end{cases}$$

计算 $x(0.1)$(用计算器).

4. 利用常微分方程

$$\begin{cases} x' = x^2 + xe^t \\ x(0) = 1 \end{cases}$$

和三阶泰勒级数方法一步计算 $x(0.01)$.

5. 考虑常微分方程

$$\begin{cases} 5tx' + x^2 = 2 \\ x(4) = 1 \end{cases}$$

利用二阶泰勒级数方法一步计算 $x(4.1)$.

6. **积分方程**是在积分中包含一个未知函数的方程. 例如，下面是一个典型的积分方程(用名字 Volterra 命名的一类积分方程)：

$$x(t) = \int_0^t \cos(s + x(s))\mathrm{d}s + e^t$$

通过对这个积分方程求微分，得到未知函数的一个等价的初值问题.

7. 若泰勒级数方法用于求解涉及微分方程

$$x' = \cos(tx)$$

的初值问题，试问 x''，x'''和 $x^{(4)}$ 的公式是什么?

8. 设 $x' = f(t, x)$，从这个方程确定 x''，x'''和 $x^{(4)}$.

计算机习题 8.2

[536]

1. 编写并测试在区间[1, 3]上求解下列带有初始条件的微分方程

$$\begin{cases} x' = x + e^t + tx \\ x(1) = 2 \end{cases}$$

的计算机程序. 利用五阶泰勒级数方法和 $h=0.01$.

2. 编写并测试求解下列初值问题

$$\begin{cases} x' = 1 + x^2 - t^3 \\ x(0) = -1 \end{cases}$$

的计算机程序. 利用四阶泰勒级数方法，h 为接近于 0.01 的二进制机器数. 在区间[0, 2]中求解. 说明解中任何特别的现象.

3. 求解初值问题的方法也可能用于计算定积分或不定积分. 例如，通过在 t-区间[0, 2]上求解初值问题.

$$\begin{cases} x' = e^{-t^2} \\ x(0) = 0 \end{cases}$$

可计算

$$\int_0^2 e^{-s^2}\,\mathrm{d}s$$

利用四阶泰勒级数方法来做此项工作. 根据误差函数

$$\mathrm{erf}(t) = \frac{2}{\sqrt{\pi}}\int_0^t e^{-s^2}\,\mathrm{d}s$$

的表，我们得到 $x(t) \approx 0.882\,081\,390\,7$.(见 Abramowitz and Stegun[1964，第 311 页].)

4. (续)利用习题 8.2.1 编写可应用于上题中的积分的六阶泰勒级数方法的计算机程序. 用 $h=0.01$ 测试代码，看看是否得到 $x(2)$的相同的值. 打印函数$\frac{1}{2}\sqrt{\pi}\mathrm{erf}(t)$的表格，从 $t=0$ 到 $t=2$，步长为 0.01.

5. 利用四阶泰勒级数方法，取 $h=0.01$，在区间[0，1.56]上求解初值问题

$$\begin{cases} x' = 1 + x^2 \\ x(0) = 0 \end{cases}$$

然后，利用 $x(1.56)$的计算值作为初值积分回到 $t=0$. 比较结果并说明发生什么情况.

6. 等式 $\arctan(x/t)=\ln\sqrt{x^2+t^2}$隐式地定义 x 为 t 的函数. 验证这个隐函数是初值问题

$$\begin{cases} x' = (t+x)/(t-x) \\ x(1) = 0 \end{cases}$$

的解. 作出函数 $x(t)$在区间[0，2]上步长为±0.01 的表格. 利用四阶泰勒级数方法.

7. 函数

$$\varphi(t) = \int_0^t \sin s^2 \mathrm{d}s$$

537

称为**菲涅尔积分**. 作出这个函数在区间[0，10]上的表格. 利用五阶泰勒级数方法，取 $h=0.1$. 如果可能，得到函数的计算机绘图.

8. (续)证明函数 $f(x)=\sin t^2$ 的导数是

$$f^{(n)}(t) = P_n(t)\sin t^2 + Q_n(t)\cos t^2$$

其中 P_n 和 Q_n 递归计算如下：

$$\begin{cases} P_0 = 1 \\ P_{n+1} = P'_n - 2tQ_n \end{cases} \qquad \begin{cases} Q_0 = 0 \\ Q_{n+1} = Q'_n + 2tP_n \end{cases}$$

利用上式生成 $n=0$ 到 6 的 P_n 和 Q_n 的值表. 修正上题的代码变为七阶的泰勒级数方法. 像在上题中那样，先用 $h=0.1$ 再用 $h=0.2$ 进行测试. 把用 $h=0.2$ 得到的 $\varphi(10)$的值和上题得到的值作比较.

9. 函数

$$x(t) = \int_0^t (1 - k\sin^2\theta)^{1/2} \mathrm{d}\theta$$

是**第 2 类椭圆积分**，其中 k 是[0，1]中的一个参数. 当 $k=1/2$ 时，利用三阶泰勒级数方法，取 $h=0.01$，在区间 $0\leqslant t\leqslant \pi/2$ 上作出这个函数的表格.

10. 利用二阶泰勒级数方法，在区间[0，1]上，取步长 $h=-0.01$，求解初值问题

$$\begin{cases} x' = -\dfrac{3t^2x + x^2}{2t^3 + 3tx} \\ x(1) = -2 \end{cases}$$

利用隐式解 $t^3x^2+tx^3+4=0$ 检验计算解. 并且验证隐式解的正确性.

11. 通过解一个适当的初值问题在区间 $-2\leqslant x\leqslant 0$ 上作出**二重对数函数**

$$f(x) = -\int_0^x \frac{\ln(1-t)}{t} \mathrm{d}t$$

的表格. (与习题 6.7.14 作比较.)

12. 积分 $\int\sqrt{1+x^3}\mathrm{d}x$ 不能用初等微积分得到. (它是一个**椭圆积分.**) 通过解一个适当的初值问题在区间 $0\leqslant x\leqslant 5$ 上作出函数

$$f(x) = \int_0^x \sqrt{1+t^3}\mathrm{d}t$$

的表格. 利用三阶泰勒级数方法，取 $h=1/64$.

13. 考虑初值问题 $x'=1-xt^{-1}$，$x(2)=2$. 证明

538

$$\begin{cases} x'' = (1-2x')t^{-1} \\ x^{(n)} = -nx^{(n-1)}t^{-1} & (n \geqslant 3) \end{cases}$$

编写用 10 阶泰勒级数方法求解这个初值问题的计算机程序．用 $h=1$ 并在区间[2，20]上测试你的程序．该问题的解析解是什么？把它与你的数值解进行比较．（见 Conte and de Boor[1980].）

14. 编写用泰勒级数方法求解初值问题

$$\begin{cases} x' = t^2 + x^2 + 2tx \\ x(0) = 7 \end{cases}$$

的计算机程序．需要求的是在区间 $-2 \leqslant t \leqslant 2$ 内的解．包括按升次直到含 h^3 的那些项．利用步长 0.01.

15. 编写计算函数

$$x(t) = \int_2^t \sin(u^3)\mathrm{d}u$$

值表的计算机程序．通过建立一个适当的初值问题来做此项工作．然后利用三阶泰勒级数方法来实现求数值解．估计所产生的表的精度．

16. 利用三阶泰勒级数方法就等价的初值问题求解习题 8.2.6 中的积分方程．利用步长 0.01 得到[0，3]上的解．

17. 利用符号操作程序，用四阶及 20 阶泰勒级数方法来求解课本中(3)式给出的初值问题．与课本中的数值结果作比较．利用各种步长 h 比较这两个泰勒级数方法．为了得到与用 $h=0.01$ 的四阶方法相同的精度，确定 20 阶方法中 h 取多大的值．

18. (续)用上题中的方法求解计算机习题 8.2.6 中的初值问题．

19. 用包括(5)式中给出的 E_4 改编伪代码．

20. 在区间 $0 \leqslant t \leqslant 1$ 上数值求解初值问题

$$\begin{cases} (x')^2 - 2tx' - x\cos t = 0 \\ x(0) = 0 \end{cases}$$

(这里的平凡解不是我们所想要的！) 若 $(x')^2$ 改成 $(x')^3$，你将会做什么？

21. 数值求解方程(6)并把它与真解比较．

22. 利用三阶泰勒级数方法在区间[0，1]上用步长 $h=0.01$ 实现求方程(7)的数值解．

8.3 龙格-库塔方法

上节中的泰勒级数方法的不足之处在于为了对它进行程序设计，需要事先作些分析．例如，如果希望对一般的问题

539

$$\begin{cases} x' = f(t,x) \\ x(t_0) = x_0 \end{cases} \tag{1}$$

用四阶泰勒级数方法，就必须逐次微分(1)式来确定 x''，x'''和 $x^{(4)}$．然后，必须对这些函数编程．

虽然龙格-库塔方法仿效泰勒级数方法使用 $f(t, x)$值的巧妙组合，但它们避免了上述困难．我们通过推导二阶龙格-库塔方法来说明这一点．

8.3.1 二阶龙格-库塔方法

我们从 $x(t+h)$的泰勒级数入手：

$$x(t+h) = x(t) + hx'(t) + \frac{h^2}{2!}x''(t) + \frac{h^3}{3!}x'''(t) + \cdots \tag{2}$$

根据微分方程，我们有

$$
\begin{aligned}
x'(t) &= f \\
x''(t) &= f_t + f_x x' = f_t + f_x f \\
x'''(t) &= f_{tt} + f_{tx} f + (f_t + f_x f) f_x + f(f_{xt} + f_{xx} f) \\
&\vdots
\end{aligned}
$$

这里下标表示偏导数，反复地利用微分法的链式法则.（2）式中的前三项现在可写成下列形式

$$
\begin{aligned}
x(t+h) &= x + hf + \frac{1}{2}h^2(f_t + ff_x) + \mathcal{O}(h^3) \\
&= x + \frac{1}{2}hf + \frac{1}{2}h[f + hf_t + hff_x] + \mathcal{O}(h^3)
\end{aligned} \tag{3}
$$

其中 x 表示 $x(t)$，f 表示 $f(t, x)$，等等. 我们借助于两个变量的泰勒级数（见 1.1 节）中的前几项可消去偏导数：

$$
f(t+h, x+hf) = f + hf_t + hff_x + \mathcal{O}(h^2)
$$

（3）式可改写成

$$
x(t+h) = x + \frac{1}{2}hf + \frac{1}{2}hf(t+h, x+hf) + \mathcal{O}(h^3)
$$

因此，步进求解的公式是

$$
x(t+h) = x(t) + \frac{h}{2}f(t,x) + \frac{h}{2}f(t+h, x+hf(t,x))
$$

或等价地，

$$
x(t+h) = x(t) + \frac{1}{2}(F_1 + F_2) \tag{4}
$$

其中

$$
\begin{cases}
F_1 = hf(t,x) \\
F_2 = hf(t+h, x+F_1)
\end{cases}
$$

540

这个公式可反复地使用，每次前进一步求解. 它被称为**二阶龙格-库塔方法**，也称为 **Heun 方法**.

一般说来，龙格-库塔公式具有形式

$$
x(t+h) = x + w_1 hf + w_2 hf(t + \alpha h, x + \beta hf) + \mathcal{O}(h^3) \tag{5}
$$

其中 w_1，w_2，α 和 β 是我们配置的参数.（5）式可借助于两个变量的泰勒级数改写成

$$
x(t+h) = x + w_1 hf + w_2 h[f + \alpha hf_t + \beta hff_x] + \mathcal{O}(h^3) \tag{6}
$$

比较（3）式和（6）式，我们看到应该强加条件：

$$
\begin{cases}
w_1 + w_2 = 1 \\
w_2 \alpha = \dfrac{1}{2} \\
w_2 \beta = \dfrac{1}{2}
\end{cases} \tag{7}
$$

一个解是 $w_1=w_2=\frac{1}{2}$，$\alpha=\beta=1$，这就是对应于(4)式中的 Heun 方法. 方程组(7)有不同于这个解的解，譬如取 $w_1=0$，$w_2=1$，$\alpha=\beta=\frac{1}{2}$得到的解. 由(5)式得到的公式称为**修正的欧拉方法**：

$$x(t+h)=x(t)+F_2$$

其中

$$\begin{cases} F_1=hf(t,x) \\ F_2=hf\left(t+\frac{1}{2}h,x+\frac{1}{2}F_1\right) \end{cases}$$

把这个方法与 8.2 节中所述的标准的欧拉方法作比较.

8.3.2 四阶龙格-库塔方法

推导高阶龙格-库塔公式是非常令人乏味的，我们将不做这些工作. 然而，公式是相当漂亮的，而且一旦它们被推导出后，编程是很容易的. 下面是经典的**四阶龙格-库塔方法**：

$$x(t+h)=x(t)+\frac{1}{6}(F_1+2F_2+2F_3+F_4) \tag{8}$$

其中

$$\begin{cases} F_1=hf(t,x) \\ F_2=hf\left(t+\frac{1}{2}h,x+\frac{1}{2}F_1\right) \\ F_3=hf\left(t+\frac{1}{2}h,x+\frac{1}{2}F_2\right) \\ F_4=hf(t+h,x+F_3) \end{cases}$$

这个方法被称为四阶方法是因为它再现了泰勒级数中按升幂直到包含涉及 h^4 的项. 所以误差是 $\mathcal{O}(h^5)$. h^5 误差项的精确表达式是可利用的.

541

例 1 给出一个利用四阶龙格-库塔方法的算法，在区间[1, 3]上，用步长 $h=1/128$ 求解下列初值问题：

$$\begin{cases} x'=t^{-2}(tx-x^2) \\ x(1)=2 \end{cases} \tag{9}$$

解 因为这是一个测量方法有效性的数值实验，所以选择一个已知解析解的问题. (9)的解为 $x(t)=\left(\frac{1}{2}+\ln t\right)^{-1}t$. 误差值由计算机程序打印.

```
input M←256; t←1.0; x←2.0; h←0.0078125
define f(t, x)=(tx−x²)/t²
define u(t)=t/(1/2+ln t)
e←|u(t)−x|
output 0, t, x, e
for k=1 to M do
```

$F_1 \leftarrow hf(t, x)$

$F_2 \leftarrow hf\left(t+\frac{1}{2}h, x+\frac{1}{2}F_1\right)$

$F_3 \leftarrow hf\left(t+\frac{1}{2}h, x+\frac{1}{2}F_2\right)$

$F_4 \leftarrow hf(t+h, x+F_3)$

$x \leftarrow x+(F_1+2F_2+2F_3+F_4)/6$

$t \leftarrow t+h$

$e \leftarrow |u(t)-x|$

output k, t, x, e

end do

基于这个算法的计算机程序的一些输出数据如下：

k	t	x	e
0	1.000 00	2.000 00	
1	1.007 81	1.984 73	1.19×10^{-7}
2	1.015 63	1.970 16	0
3	1.023 44	1.956 23	0
4	1.031 25	1.942 93	1.19×10^{-7}
5	1.039 06	1.930 20	1.19×10^{-7}
6	1.046 88	1.918 02	0
7	1.054 69	1.906 37	1.19×10^{-7}
8	1.062 50	1.895 21	1.19×10^{-7}
9	1.070 31	1.884 52	0
⋮	⋮	⋮	⋮
249	2.945 31	1.863 87	7.15×10^{-7}
250	2.953 13	1.865 69	5.96×10^{-7}
251	2.960 94	1.867 50	7.15×10^{-7}
252	2.968 75	1.869 32	5.96×10^{-7}
253	2.976 56	1.871 15	4.77×10^{-7}
254	2.984 38	1.872 97	5.96×10^{-7}
255	2.992 19	1.874 80	5.96×10^{-7}
256	3.000 00	1.876 63	5.96×10^{-7}

542

■

8.3.3 误差

我们现在转到讨论龙格-库塔方法中的**截断误差**. 粗略地讲，局部截断误差就是在每一步中产生的误差，它仅仅是由于我们的方法不能考虑泰勒级数的所有项所引起的. 这个误差是不可避免的，并且即使计算以无限精度执行的话，误差也会出现. 在四阶龙格-库塔方法的情况中，公式涉及一个$\mathcal{O}(h^5)$的局部截断误差. 当按升幂直到h^4的项包含在泰勒公式中时，这个

误差的阶正好对应于泰勒级数方法中的误差.

在四阶龙格-库塔方法的第1步，$x(t_0+h)$的值是由算法计算出来的. 另一方面，存在一个我们不知道的正确解 $x^*(t_0+h)$. 由定义，在该步中的局部截断误差是

$$x^*(t_0+h)-x(t_0+h)$$

龙格-库塔算法的理论指出这个截断误差对小的 h 值具有类似 Ch^5 的性态. 这里 C 是一个与 h 无关但与 t_0 和函数 x^* 有关的数. 为估计 Ch^5，我们假定当 t 从 t_0 变到 t_0+h 时 C 不变. 设 v 是近似解在 t_0+h 上的值，它是从 t_0 出发取长度为 h 的一步得到的. 设 u 是在 t_0+h 的近似解，它是从 t_0 出发取长度为 $h/2$ 的二步得到的. 这两者都是可计算的. 利用所作的假定，我们有

$$x^*(t_0+h)=v+Ch^5$$

$$x^*(t_0+h)=u+2C(h/2)^5$$

把这两式相减，得到

$$\text{局部截断误差}=Ch^5=(u-v)/(1-2^{-4})$$

因此，局部截断误差近似等于 $u-v$.

在龙格-库塔方法的计算机实现中，为保证近似的截断误差处于特定的容限之下，偶尔可以通过计算 $|u-v|$ 来控制. 若它不行，则可减小步长(通常减半)来改善局部截断误差. 另一方面，若局部截断误差远离容许的阈值，则步长可以加倍.

543

正如我们在二阶龙格-库塔方法的推导中看到过的那样，必须选择若干参数. 在建立高阶龙格-库塔方法中出现类似的选择过程. 因此，各阶龙格-库塔方法不仅仅有一个而且有一族方法. 正如下表所示的那样，需要的函数赋值数比龙格-库塔方法阶数增加得更迅速.

函数赋值数	1	2	3	4	5	6	7	8
龙格-库塔方法的最大阶	1	2	3	4	4	5	6	6

遗憾的是，这就使得高阶龙格-库塔方法比经典的四阶方法吸引力小，因为使用它们所需的计算量更多.

8.3.4 自适应龙格-库塔-费尔贝格方法

为了尝试在龙格-库塔方法中设计一个自动调整步长的方法，Fehlberg[1969]转向具有5个函数赋值的四阶方法和具有6个函数赋值的五阶方法. 初看起来这似乎是没有指望的，因为他的四阶方法比经典的四阶方法需要更多的函数赋值. 而且，这两个方法一起总共需要7个函数赋值. 然而，Fehlberg能在这些方法中选择参数从而得到具有相同函数赋值点的不同阶数的两个公式. 因此，只需要6个函数赋值，便得到一个基于步长控制的局部截断误差的估计. 所得到的**龙格-库塔-费尔贝格方法**是五阶的，并且利用四阶和五阶的两个公式. 这些公式给出解的不同的近似值，我们用 $x(t+h)$和 $\bar{x}(t+h)$来表示它们：

$$x(t+h)=x(t)+\sum_{i=1}^{6}a_iF_i \tag{10}$$

$$\bar{x}(t+h)=x(t)+\sum_{i=1}^{6}b_iF_i \tag{11}$$

量 F_i 由下列形式的公式算出：

$$F_i = hf\left(t + c_i h, x + \sum_{j=1}^{i-1} d_{ij} F_j\right) \qquad (1 \leqslant i \leqslant 6) \tag{12}$$

公式(10)是五阶的，而公式(11)是四阶的．差 $e = x(t+h) - \bar{x}(t+h)$ 可解释为对应于不太精确的公式(11)的局部截断误差的估计．因此，e 可用于控制步长．因为公式(10)更精确，所以它用于提供算法中的输出．注意

544

$$e = x(t+h) - \bar{x}(t+h) = \sum_{i=1}^{6} (a_i - b_i) F_i \tag{13}$$

系数的值在下表中给出：

i	a_i	$a_i - b_i$	c_i	d_{ij}
1	$\frac{16}{135}$	$\frac{1}{360}$	0	0
2	0	0	$\frac{1}{4}$	$\frac{1}{4}$
3	$\frac{6\,656}{12\,825}$	$-\frac{128}{4\,275}$	$\frac{3}{8}$	$\frac{3}{32}, \frac{9}{32}$
4	$\frac{28\,561}{56\,430}$	$-\frac{2\,197}{75\,240}$	$\frac{12}{13}$	$\frac{1\,932}{2\,197}, -\frac{7\,200}{2\,197}, \frac{7\,296}{2\,197}$
5	$-\frac{9}{50}$	$\frac{1}{50}$	1	$\frac{439}{216}, -8, \frac{3\,680}{513}, -\frac{845}{4\,104}$
6	$\frac{2}{55}$	$\frac{2}{55}$	$\frac{1}{2}$	$-\frac{8}{27}, 2, -\frac{3\,544}{2\,565}, \frac{1\,859}{4\,104}, -\frac{11}{40}$

利用前面公式的一个自适应例程试图保持局部截断误差 e 的值在某个预先给定的容限 δ 之下．在某个区间 $[a, b]$ 中求解，并且初值 $x(a) = \alpha$ 也给出．步数的上界 M 也预先给定．注意，因为四阶方法中的局部截断误差被用于控制步长而数值解实际上是用五阶公式计算的，所以该方法是保守的．下面给出执行这个方法的一个算法．因为用 $x(t) - \bar{x}(t)$ 估计的局部截断误差为 Ch^5，所以当步长加倍导致 $C(2h)^5 < \delta/4$ 时，加倍步长看起来是合理的．因此，当局部截断误差明显小于 $\delta/128$ 时，我们将步长 h 加倍．

另一个通用的每步控制误差的公式是：

$$h \leftarrow 0.9h[\delta / |e|]^{1/(1+p)}$$

其中 p 是一对龙格-库塔公式中第 1 个公式的阶（例如，见 Hull，Enright，Fellen and Sedgwick[1972]以及 Shampine，Watts and Davenport[1976].）同样的公式既被用来增加步长也被用来减短步长．在一种情况下，重算不合适的步长，而在其他情况下，则建立一个新的步长．

下面是**自适应龙格-库塔-费尔贝格算法：**

```
input a, α, b, h, δ, M
t←a; x←α
iflag←1
while k<M do
    d←b−t
    if |d| ≤ |h| then
        iflag←0
```

```
            h←d
        end if
        s←t
        y←x
        compute F1, F2, …, F6  [Equation (12)]
        compute x  [Equation (10)]
        compute e  [Equation (13)]
          t←t+h
          output k, t, x, e
        if iflag=0 then stop
        if |e| ≥ δ then
                t←s
                h←h/2
                x←y
                k←k−1
         else
              if |e| < δ/128 then h←2h
         end if
    end do
```

545

嵌入龙格-库塔方法是由一对 p 阶和 q 阶的公式组成的(通常 $q=p+1$)，这对公式在同样的点上计算函数值. 一般说来，它们提供一个求解非刚性初值问题的有效方法. 近年来已推导出许多高阶嵌入龙格-库塔公式. 在习题中给出了这些公式的某些例子. 虽然带有复杂系数的高阶龙格-库塔方法已被推导出来，但是除了自适应格式外，经典的四阶公式仍是最流行的. 关于龙格-库塔方法的附加信息可在为数众多的原始资料中找到——例如，Butcher[1987]、Fehlberg[1969]、Gear[1971]、Jackson, Enright and Hull[1978]、Prince and Dormand[1981]、Shampine and Gordon[1975]、Thomas[1986]以及 Verner[1978].

习题 8.3

1. 写出 $\alpha=2/3$ 时的二阶龙格-库塔公式.
2. 若利用龙格-库塔方法求解涉及微分方程 $x+2tx'+xx'=3$ 的初值问题，试问必须对什么函数进行编程?
3. 证明龙格-库塔公式(8)在 $f(t, x)$与 x 无关的特殊情况中是四阶的. 说明此时的龙格-库塔公式等价于辛普森法则. (见 7.2 节中的(6)式.)
4. 利用步长 h 和 $h/2$ 对欧拉方法执行理查森外推推导修正的欧拉方法.
$$x(t+h)=x(t)+hf\left(t+\frac{1}{2}h, x(t)+\frac{1}{2}hf(t,x(t))\right)$$
提示：假定误差项是 Ch^2.
5. 推导三阶龙格-库塔公式
$$x(t+h)=x(t)+\frac{1}{9}(2F_1+3F_2+4F_3)$$
其中
$$\begin{cases} F_1=hf(t,x) \\ F_2=hf(t+\frac{1}{2}h, x+\frac{1}{2}F_1) \\ F_3=hf(t+\frac{3}{4}h, x+\frac{3}{4}F_2) \end{cases}$$

546

说明它与微分方程 $x'=x+t$ 的三阶泰勒级数方法一致.

6. 证明当四阶龙格-库塔方法应用于问题 $x'=\lambda x$ 时，步进求解的公式将是

$$x(t+h)=\left[1+h\lambda+\frac{1}{2}h^2\lambda^2+\frac{1}{6}h^3\lambda^3+\frac{1}{24}h^4\lambda^4\right]x(t)$$

7. (续)证明上题中局部截断误差是 $\mathcal{O}(h^5)$.

计算机习题 8.3

1. 编写在区间 $t_0\leqslant t\leqslant t_m$ 或 $t_m\leqslant t\leqslant t_0$ 上求解初值问题 $x'=f(t, x)$，$x(t_0)=x_0$ 的计算机程序. 使用四阶龙格-库塔方法. 对下列例子测试这个程序：

$$\begin{cases}(\mathrm{e}^t+1)x'+x\mathrm{e}^t-x=0\\ x(0)=3\end{cases}$$

确定解析解并把它与区间 $-2\leqslant t\leqslant 0$ 上的计算解作比较. 利用 $h=-0.01$.

2. 利用四阶龙格-库塔方法和各种 λ 的值，譬如，5，-5 或 -10，数值求解下列初值问题：

$$\begin{cases}x'=\lambda x+\cos t-\lambda\sin t\\ x(0)=0\end{cases}$$

在区间[0，5]上比较数值解和解析解. 利用步长 $h=0.01$. 试问 λ 对数值准确性有什么影响？

3. 对文中描述的龙格-库塔-费尔贝格算法编程且测试. 测试时，使用计算机习题 8.2.7、8.2.11、8.2.15 中的例子.

4. 编写并执行求解初值问题

$$\begin{aligned}x'&=\mathrm{e}^{xt}+\cos(x-t)\\ x(1)&=3\end{aligned}$$

的程序. 利用四阶龙格-库塔公式，取 $h=0.01$. 要求恰好在解上溢之前停止计算.

5. 利用自适应龙格-库塔-费尔贝格方法在区间[0，2]上求解 $x'=x^2$，$x(0)=1$. 证明真解是 $x(t)=1/(1-t)$. 在不连续点 $t=1$ 附近，算法中发生了什么情况？

6. 在数值上将下列四阶龙格-库塔-Gill 方法与经典的龙格-库塔方法进行比较.

$$x(t+h)=x(t)+\frac{1}{6}[F_1+(2-\sqrt{2})F_2+(2+\sqrt{2})F_3+F_4]$$

其中

$$\begin{cases}F_1=hf(t,x)\\ F_2=hf\left(t+\frac{1}{2}h,x+\frac{1}{2}F_1\right)\\ F_3=hf\left(t+\frac{1}{2}h,x+\frac{1}{2}(\sqrt{2}-1)F_1+\frac{1}{2}(2-\sqrt{2})F_2\right)\\ F_4=hf\left(t+h,x-\frac{1}{2}\sqrt{2}F_2+\frac{1}{2}(2+\sqrt{2})F_3\right)\end{cases}$$

547

7. 对一个具有已知解的问题，在数值上将下列五阶龙格-库塔方法与经典的龙格-库塔方法进行比较.

$$x(t+h)=7x(t)+\frac{1}{24}F_1+\frac{5}{48}F_4+\frac{27}{56}F_5+\frac{125}{336}F_6$$

其中

$$\begin{cases}F_1=hf(t,x)\\ F_2=hf\left(t+\frac{1}{2}h,x+\frac{1}{2}F_1\right)\\ F_3=hf\left(t+\frac{1}{2}h,x+\frac{1}{4}F_1+\frac{1}{4}F_2\right)\\ F_4=hf(t+h,x-F_2+2F_3)\\ F_5=hf\left(t+\frac{2}{3}h,x+\frac{7}{27}F_1+\frac{10}{27}F_2+\frac{1}{27}F_4\right)\\ F_6=hf\left(t+\frac{1}{5}h,x+\frac{28}{625}F_1-\frac{1}{5}F_2+\frac{546}{625}F_3+\frac{54}{625}F_4-\frac{378}{625}F_5\right)\end{cases}$$

8. 由下表中的系数给出的是五阶**龙格-库塔-Merson 方法**. 编写且测试基于这个方法的自适应程序.

i	a_i	a_i-b_i	c_i	d_{ij}
1	$\frac{1}{6}$	$\frac{1}{15}$	0	0
2	0	0	$\frac{1}{3}$	$\frac{1}{3}$
3	0	$-\frac{3}{10}$	$\frac{1}{3}$	$\frac{1}{6},\frac{1}{6}$
4	$\frac{2}{3}$	$\frac{4}{15}$	$\frac{1}{2}$	$\frac{1}{8},0,\frac{3}{8}$
5	$\frac{1}{6}$	$-\frac{1}{30}$	1	$\frac{1}{2},0,-\frac{3}{2},2$

9. 由下表中的系数给出的是**五阶龙格-库塔-Verner 方法**(见 Verner[1978]). 编写且测试基于这个方法的自适应程序.

i	a_i	a_i-b_i	c_i	d_{ij}
1	$\frac{3}{80}$	$\frac{33}{640}$	0	0
2	0	0	$\frac{1}{18}$	$\frac{1}{18}$
3	$\frac{4}{25}$	$-\frac{132}{325}$	$\frac{1}{6}$	$-\frac{1}{12},\frac{1}{4}$
4	$\frac{243}{1\,120}$	$\frac{891}{2\,240}$	$\frac{2}{9}$	$-\frac{2}{81},\frac{4}{27},\frac{8}{81}$
5	$\frac{77}{160}$	$-\frac{33}{320}$	$\frac{2}{3}$	$\frac{40}{33},-\frac{4}{11},-\frac{56}{11},\frac{54}{11}$
6	$\frac{73}{700}$	$-\frac{73}{700}$	1	$-\frac{369}{73},\frac{72}{73},\frac{5\,380}{219},-\frac{12\,285}{584},\frac{2\,695}{1\,752}$
7	0	$\frac{891}{8\,320}$	$\frac{8}{9}$	$-\frac{8\,716}{891},\frac{656}{297},\frac{39\,520}{891},-\frac{416}{11},\frac{52}{27}$
8	0	$\frac{2}{35}$	1	$\frac{3\,015}{256},-\frac{9}{4},-\frac{4\,219}{78},\frac{5\,985}{128},-\frac{539}{384},\frac{693}{3\,328}$

10. 由下表中的系数给出的是六阶和五阶的一对**嵌入龙格-库塔公式**(见 Prince and Dormand[1981]). 编写且测试利用它们的自适应程序.

i	a_i	b_i	c_i	d_{ij}
1	$\frac{821}{10\,800}$	$\frac{61}{864}$	0	0
2	0	0	$\frac{1}{10}$	$\frac{1}{10}$
3	$\frac{19\,683}{71\,825}$	$\frac{98\,415}{321\,776}$	$\frac{2}{9}$	$-\frac{2}{81},\frac{20}{81}$
4	$\frac{175\,273}{912\,600}$	$\frac{16\,807}{146\,016}$	$\frac{3}{7}$	$\frac{615}{1\,372},-\frac{270}{343},\frac{1\,053}{1\,372}$
5	$\frac{395}{3\,672}$	$\frac{1\,375}{7\,344}$	$\frac{3}{5}$	$\frac{3\,243}{5\,500},-\frac{54}{55},\frac{50\,949}{71\,500},\frac{4\,998}{17\,875}$
6	$\frac{785}{2\,704}$	$\frac{1\,375}{5\,408}$	$\frac{4}{5}$	$-\frac{26\,492}{37\,125},\frac{72}{55},\frac{2\,808}{23\,375},-\frac{24\,206}{37\,125},\frac{338}{459}$
7	$\frac{3}{50}$	$-\frac{37}{1\,120}$	1	$\frac{5\,561}{2\,376},-\frac{35}{11},-\frac{24\,117}{31\,603},\frac{899\,983}{200\,772},-\frac{5\,225}{1\,836},\frac{3\,925}{4\,056}$
8	0	$\frac{1}{10}$	1	$\frac{465\,467}{266\,112},-\frac{2\,945}{1\,232},-\frac{5\,610\,201}{14\,158\,144},\frac{10\,513\,573}{3\,212\,352},-\frac{424\,325}{205\,632},\frac{376\,225}{454\,272},0$

548

11. 在区间[0, 1]上考察具有各种蜕变常数 k 的初值问题 $x'=-kx$, $x(0)=1$. 证明解析解是 $x(t)=e^{-kt}$. 对 $k=5$,利用各种步长 h 的值，譬如，0.2，0.4，0.6，0.8 和 1.0，比较欧拉方法和龙格-库塔-Gill 方法(计算机习题 8.3.6)的性态. 利用 $k=25$ 重复上面的工作. 为了正确地求解这个问题，欧拉方法必须满足条件 $0\leqslant hk\leqslant 2$，而龙格-库塔-Gill 方法需要满足 $0\leqslant hk\leqslant 2.8$. 这些是这个问题利用这两个方法的**稳定性区域**. 对于特殊的问题和方法，我们可以调整步长 h 使其处于稳定性区域中，也可以另外使用具有较大稳定性区域的高阶方法. 正如 Thomas[1986]讨论的那样，对于大的 k 值，这些方法变得难以操作.

8.4 多步法

求解初值问题的泰勒级数方法和龙格-库塔方法是**单步法**，因为当解从 t 前进到 $t+h$ 时，他们没有使用任何先前的 $x(t)$ 值的知识. 若 t_0, t_1, t_2, …, t_i 是沿 t 轴的步长，则 x_{i+1}($x(t_{i+1})$的近似值)不使用近似值 x_{i-1}, x_{i-2}, …, x_0 的信息，而仅仅与 x_i 有关.

如果在每一步利用解的某些先前的值，则可以设计更有效的方法. 其中包含的原理如下：在数值上求解初值问题

$$\begin{cases}x' = f(t,x)\\ x(t_0) = x_0\end{cases} \tag{1}$$

我们给定了 t 轴上的步长 t_0, t_1, t_2, …, t_n(它们通常不是等距的). 若真解用 $x(t)$表示，则积分方程(1)给出

$$\int_{t_n}^{t_{n+1}} x'(t)\mathrm{d}t = x(t_{n+1}) - x(t_n)$$

于是

$$x(t_{n+1}) = x(t_n) + \int_{t_n}^{t_{n+1}} f(t,x(t))\mathrm{d}t \tag{2}$$

右边的积分可用数值积分格式逼近，结果是一个逐步地生成近似解的公式.

8.4.1 亚当斯-巴什福思公式

549

假设得到的公式具有下列类型：

$$x_{n+1} = x_n + af_n + bf_{n-1} + cf_{n-2} + \cdots \tag{3}$$

其中 f_i 表示 $f(t_i, x_i)$. 这类式子称为**亚当斯-巴什福思公式**. 下面是基于等距点 $t_i = x_0 + ih$($1\leqslant i\leqslant n$)的**五阶亚当斯-巴什福思公式**：

$$x_{n+1} = x_n + \frac{h}{720}[1\,901 f_n - 2\,774 f_{n-1} + 2\,616 f_{n-2} - 1\,274 f_{n-3} + 251 f_{n-4}] \tag{4}$$

这些系数是怎样确定的呢？我们首先逼近(2)式中的积分为

$$\int_{t_n}^{t_{n+1}} f(t,x(t))\mathrm{d}t \approx h[Af_n + Bf_{n-1} + Cf_{n-2} + Df_{n-3} + Ef_{n-4}] \tag{5}$$

系数 A, B, C, D 和 E 是这样确定的：当被积函数是次数$\leqslant 4$ 的多项式时，我们要求(5)式精确成立. 由习题 8.3.6，如图 8-1 所示，不失一般性，假定 $t_n=0$, $h=1$.

图 8-1 五阶亚当斯-巴什福思公式(4)中包含的点

我们取下列5个多项式作为Π_4的基：

$$\begin{aligned}p_0(t)&=1\\p_1(t)&=t\\p_2(t)&=t(t+1)\\p_3(t)&=t(t+1)(t+2)\\p_4(t)&=t(t+1)(t+2)(t+3)\end{aligned}$$

将这些多项式代入下式：

$$\int_0^1 p_n(t)\mathrm{d}t=Ap_n(0)+Bp_n(-1)+Cp_n(-2)+Dp_n(-3)+Ep_n(-4)$$

550 我们就得到确定系数A，B，C，D和E的5个方程．这个方程组是

$$\begin{cases}A+B+C+D+E=1\\-B-2C-3D-4E=1/2\\2C+6D+12E=5/6\\-6D-24E=9/4\\24E=251/30\end{cases}\tag{6}$$

通过向后回代就得到公式(4)中的系数．

刚才说明的过程称为**待定系数法**．原则上，它可用来得到高阶和各种各样其他情况的类似公式．（见7.2节．）

另外，因为$x'=f$，所以取$x=p_n$并且利用$f=p'_n$可以直接从(4)式得到这些未知系数的值．我们可以得到类似于(6)的方程组．

8.4.2 亚当斯-莫尔顿公式

在数值实践中，很少使用亚当斯-巴什福思公式本身．它们与其他公式联合起来使用以提高精度．要了解这一点，我们回到(2)式，并假定使用一个包含f_{n+1}的数值积分公式．于是，(3)式具有形式

$$x_{n+1}=x_n+af_{n+1}+bf_n+cf_{n-1}+\cdots\tag{7}$$

下面是一个这种类型的公式，称为**五阶亚当斯-莫尔顿公式**：

$$x_{n+1}=x_n+\frac{h}{720}[251f_{n+1}+646f_n-264f_{n-1}+106f_{n-2}-19f_{n-3}]\tag{8}$$

这个公式也可利用待定系数法导出．注意，它不能直接地用于步进求解，因为x_{n+1}同时出现在式子的两边！记得f_i表示$f(t_i,x_i)$，所以包含f_{n+1}的项仅仅在x_{n+1}知道后才能计算．然而，一个称为**预估-校正方法**的合适的算法，使用亚当斯-巴什福思公式(4)**预估**x_{n+1}的试验值x^*_{n+1}，然后使用亚当斯-莫尔顿公式(8)计算x_{n+1}的校正值．所以在公式(8)中，我们利用公式(4)得到的预估值x^*_{n+1}赋值f_{n+1}为$f(t_{n+1},x^*_{n+1})$．

在使用这个预估-校正方法中，因为开始仅仅知道x_0，所以必须利用一个特殊的过程去启动方法．当然，龙格-库塔方法是得到x_1，x_2，x_3，x_4的理想方法．通常同阶的公式被一起使用．所以，如同计算机习题8.3.8～8.3.10中的那些方法那样，五阶龙格-库塔方法可以与亚当斯-巴什福思公式(4)和亚当斯-莫尔顿公式(8)组合在一起使用．

还有另一些方法可用来得到(8)式中 x_{n+1} 的值. 归根结底，(8)式说明 x_{n+1} 是某个映射的不动点，即映射由

$$\phi(z)=\frac{251}{720}h\,f(t_{n+1},z)+C$$

551

定义，其中 C 是由(8)式中所有其他项组成的. 所以泛函迭代法暗示它本身作为计算 x_{n+1} 的一个手段. 正如在 3.4 节中所讨论的那样，泛函迭代的理论告诉我们在适当的假设下，由

$$z_{k+1}=\phi(z_k)\qquad\qquad(k\geqslant 0)$$

确定的序列将收敛于 ϕ 的不动点. 特别地，习题 3.4.31 可在这里使用. 若 ξ 是 ϕ 的不动点(我们寻找的 x_{n+1} 的值)，则迭代应该从中心在 ξ 的区间中的一个点 z_0 出发，其中 $|\phi'(z)|<1$. 有必要假定 ϕ' 连续的. 在所探讨的情况中，

$$\phi'(z)=\frac{251}{720}h\,\frac{\partial f(t_{n+1},z)}{\partial z}$$

我们通过减小步长 h 使得上式尽可能小. 实际上，为了求 x_{n+1} 的值，在这个迭代中只需要 1 步或 2 步.

8.4.3 线性多步法的分析

在本节的其余部分，我们着手讨论一般的线性多步法理论. 这种方法的形式是

$$a_kx_n+a_{k-1}x_{n-1}+\cdots+a_0x_{n-k}=h[b_kf_n+b_{k-1}f_{n-1}+\cdots+b_0f_{n-k}]\tag{9}$$

这称为 **k 步方法**. 系数 a_i 和 b_i 已知. 如前面一样，x_i 表示解在 t_i 上的近似，$t_i=t_0+ih$. 字母 f_i 表示 $f(t_i, x_i)$. 假定 x_0，x_1，x_2，…，x_{n-1} 已知，用(9)式计算 x_n. 因此，我们可假定 $a_k\neq 0$. 系数 b_k 可为 0. 若 $b_k=0$，则称方法为**显式的**，此时 x_k 可从(9)式以初等方式直接算出. 然而，若 $b_k\neq 0$，则在右端 f_n 项中包含未知数 x_n，(9)式隐式地确定 x_n. 于是这个方法称为**隐式的**.

微分方程数值解的精度很大程度上是由使用的算法的阶确定的. 阶表明方法所模拟的泰勒级数解中有多少项. 例如，(4)式中的亚当斯-巴什福思方法被说成是五阶的，因为它近似地产生同带有 h，h^2，h^3，h^4 和 h^5 的泰勒级数方法相同的精度. 于是，在利用(4)式求解的每一步中误差可以期望是 $\mathcal{O}(h^6)$. 阶的直观概念在下面的讨论中会更精确.

对应于(9)式中的多步法，我们定义一个线性泛函 L 为

$$Lx=\sum_{i=0}^{k}[a_ix(ih)-hb_ix'(ih)]a\tag{10}$$

这里为了简化记号设 $k=n$，并假定(9)式中的第一个值在 $t=0$ 而不是在 $t=n-k$ 开始. 运算 Lx 可应用于任何可微的函数 x. 然而，在下面的分析中，假定 x 用它在 $t=0$ 的泰勒级数表示. 利用关于 x 的泰勒级数，可以把 L 表示为下列形式：

552

$$Lx=d_0x(0)+d_1hx'(0)+d_2h^2x''(0)+\cdots\tag{11}$$

为计算(11)式中的系数 d_i，我们写出 x 和 x' 的泰勒级数：

$$x(ih)=\sum_{j=0}^{\infty}\frac{(ih)^j}{j\,!}x^{(j)}(0)$$

$$x'(ih)=\sum_{j=0}^{\infty}\frac{(ih)^j}{j\,!}x^{(j+1)}(0)$$

然后把这些级数代入到(10)式中．按 h 的幂重新整理结果，我们得到一个形如(11)的式子，这些 d_i 的值为：

$$\begin{cases} d_0 = \sum_{i=0}^{k} a_i \\ d_1 = \sum_{i=0}^{k} (ia_i - b_i) \\ d_2 = \sum_{i=0}^{k} \left(\frac{1}{2}i^2 a_i - ib_i\right) \\ \vdots \\ d_j = \sum_{i=0}^{k} \left(\frac{i^j}{j\,!}a_i - \frac{i^{j-1}}{(j-1)\,!}b_i\right) \qquad (j \geqslant 1) \end{cases} \tag{12}$$

当然，这里我们使用 $0!=1$ 和 $i^0=1$.

定理1(*多步法性质定理*) *多步法*(9)*的下列三个性质等价*：

1. $d_0=d_1=\cdots=d_m=0$.
2. *对每个次数* $\leqslant m$ *的多项式*，$Lp=0$.
3. *对一切* $x\in C^{m+1}$，Lx *是* $\mathcal{O}(h^{m+1})$.

证明 若性质1成立，则(11)式具有形式

$$Lx = d_{m+1}h^{m+1}x^{(m+1)}(0) + \cdots \tag{13}$$

若 x 是次数 $\leqslant m$ 的多项式，则对一切 $j>m$，$x^{(j)}(t)=0$，因此从(13)式可得 $Lx=0$．故性质1推出性质2.

[553] 假定性质2成立．若 $x\in C^{m+1}$，则由泰勒定理我们可记 $x=p+r$，其中 p 是一个次数 $\leqslant m$ 的多项式，而 r 是一个函数，它的前 m 阶导数在0点为零．因为 $Lp=0$，所以(11)式给出

$$Lx = Lr = d_{m+1}h^{m+1}r^{(m+1)}(0) = \mathcal{O}(h^{m+1})$$

故性质2推出性质3．最后，假定性质3成立．则在(11)式中，必有条件 $d_0=d_1=\cdots=d_m=0$.因此，性质3推出性质1. ■

(9)式中的多步法的阶是唯一使得

$$d_0 = d_1 = \cdots = d_m = 0 \neq d_{m+1}$$

成立的自然数 m.

例1 由下列式子描述的方法的阶是多少？

$$x_n - x_{n-2} = \frac{1}{3}h(f_n + 4f_{n-1} + f_{n-2})$$

解 向量 (a_0,a_1,a_2) 是 $(-1,0,1)$，而向量 (b_0,b_1,b_2) 是 $\left(\frac{1}{3},\frac{4}{3},\frac{1}{3}\right)$．因此，$d_i$ 是

$$d_0 = a_0 + a_1 + a_2 = 0$$

$$d_1 = -b_0 + (a_1 - b_1) + (2a_2 - b_2) = 0$$

$$d_2 = \left(\frac{1}{2}a_1 - b_1\right) + (2a_2 - 2b_2) = 0$$

$$d_3 = \left(\frac{1}{6}a_1 - \frac{1}{2}b_1\right) + \left(\frac{4}{3}a_2 - 2b_2\right) = 0$$

$$d_4 = \left(\frac{1}{24}a_1 - \frac{1}{6}b_1\right) + \left(\frac{2}{3}a_2 - \frac{4}{3}b_2\right) = 0$$

$$d_5 = \left(\frac{1}{120}a_1 - \frac{1}{24}b_1\right) + \left(\frac{4}{15}a_2 - \frac{2}{3}b_2\right) = -\frac{1}{90}$$

所以方法的阶是 4. ■

如果其他特性不相上下的话，我们可能更喜欢高阶方法而不是低阶方法. 如果要求产生一个 $2k$ 阶的形如(9)式的 k 步方法，我们简单地写出 $2k+1$ 个等式：

$$d_0 = d_1 = \cdots = d_{2k} = 0$$

由(12)式知，这是一个 $2k+2$ 个未知数 a_i 和 $b_i(1 \leqslant i \leqslant k)$ 的 $2k+1$ 个齐次线性方程的方程组. 由初等线性代数知，这个方程组有非平凡解；Dahlquist[1956]证明了对于 $a_k \neq 0$(在方法中这是必要的)，存在一个解. 然而除了必须考虑多步法的阶之外，我们将在 8.5 节中看到多步法的某些特征，其中首要的是该节中定义的稳定性. Dahlquist 也证明了(9)式中的一个稳定的 k 步法不能有大于 $k+2$ 的阶.

554

习题 8.4

1. 导出一阶(一步)亚当斯-莫尔顿公式并验证它等价于梯形法则. (见 7.2 节.)
2. 验证(6)式中的方程组的正确性.
3. 利用待定系数法推导出(8)式.
4. 利用待定系数法推导出**四阶亚当斯-巴什福思公式**

$$x_{n+1} = x_n + \frac{h}{24}[55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}]$$

5. 推导出**四阶亚当斯-莫尔顿公式**

$$x_{n+1} = x_n + \frac{h}{24}[9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}]$$

6. 证明：若 Π_m 的每个元素式由公式

$$\int_0^1 f(x)\mathrm{d}x \approx \sum_{i=-n}^{n} A_i f(i)$$

正确地求积，则公式

$$\int_{t_0}^{t_0+h} f(x)\mathrm{d}x \approx h\sum_{i=-n}^{n} A_i f(t_0 + ih)$$

同样成立.

7. 推导出**二阶亚当斯-巴什福思公式**

$$x_{n+1} = x_n + h\left[\frac{3}{2}f_n - \frac{1}{2}f_{n-1}\right]$$

8. a. 利用待定系数法确定下列亚当斯-巴什福思公式中的 A 和 B. (不改成相应的数值积分公式.)

$$x_{n+1} = x_n + h[Af_n + Bf_{n-1}]$$

   b. 通过把它改成相应的数值积分公式，重复 a.

   c. 通过基于结点 x_n 和 x_{n-1} 的牛顿插值多项式的求积，重复 a.

9. a. 利用待定系数法求隐式二阶亚当斯-莫尔顿公式

$$x_{n+1} = x_n + h[Af_{n+1} + Bf_n]$$

中的 A 和 B.

b. 利用数值积分公式推导出该公式.

c. 利用插值公式推导出该公式.

10. a. 利用待定系数法推导出下列形式的多步法公式

$$x_{n+1} = x_n + h[Af_{n+1} + Bf_n + Cf_{n-1}]$$

b. 对

$$x_{n+1} = x_n + h[Af_n + Bf_{n-1} + Cf_{n-2}]$$

重复 a.

11. 为数值求解常微分方程 $x' = f$，推导出一个基于辛普森法则(包含等距结点 x_{n-1}，x_n，x_{n+1})的隐式多步法公式.

555

12. 已知公式

$$x_{n+1} = (1-A)x_n + Ax_{n-1} + \frac{h}{12}[(5-A)x'_{n+1} + 8(1+A)x'_n + (5A-1)x'_{n-1}]$$

对一切 A 的次数小于或等于 m 的多项式精确成立. 确定 A 使它对一切 $m+1$ 次多项式精确成立. 求 A 和 m.

13. 计算下列形式的多步法中的系数

$$x_{n+1} = x_n + h[Af_n + Bf_{n-2} + Cf_{n-4}]$$

当右端具有形式 $f(t, x) = a + bt + ct^2$ 时，公式应正确地对方程 $x' = f(t, x)$ 积分.

14. 求形如(9)式的四阶方法，取 $k=2$.

15. 证明：(9)式的多步法为 m 阶当且仅当对于 $0 \leqslant i \leqslant m$，$Lp_i = 0$，而 $Lp_{m+1} \neq 0$，其中 $p_i(t) = t^i$，而 L 如(10)式中给出的那样.

16. (续)根据上题中的证明过程，确定下列方法的阶：

$$x_n = x_{n-2} + 2hf_{n-1}$$

17. 通过计算 d_0，d_1，…确定下列方法的阶：

$$x_n = x_{n-3} + \frac{3}{8}h[f_n + 3f_{n-1} + 3f_{n-2} + f_{n-3}]$$

18. 已知 $a_0, a_1, \cdots, a_m$，且假定 $\sum_{i=0}^{k} a_i = 0$. 试问一定存在 $b_0, b_1, \cdots, b_m$ 使得具有这些系数的多步方法的阶至少为 m 吗？(需要一个定理或一个例子.)

19. 证明：对欧拉方法，$d_0 = d_1 = 0$，$d_j = 1/j!$，$j \geqslant 2$.

计算机习题 8.4

1. 编写且测试五阶亚当斯-巴什福思-莫尔顿方法的子程序或过程(见计算机习题 8.3.7). 只保留 5 个最近的 (t_i, x_i) 的值. 打印语句应包含在子程序中.

2. 编写一个计算机程序访问上题中的子程序并且在区间[−1, 1]上求解初值问题

$$\begin{cases} x' = (t - e^{-t})/(x + e^x) \\ x(0) = 0 \end{cases}$$

利用 $h = 1/238$. 验证(在解析上)真解是由等式 $x^2 - t^2 + 2e^x - 2e^{-t} = 0$ 隐式地给出. 利用程序中的这个等式检验计算解. 利用龙格-库塔方法得到初值.

3. 用 $h=0.25$ 和四阶亚当斯-巴什福思-莫尔顿方法(习题 8.4.4～8.4.5)和四阶龙格-库塔方法计算

$$\begin{cases} y' = -2xy^2 \\ y(0) = 1 \end{cases}$$

在 $x=1.0$ 上的解. 在 0.25, 0.5, 0.75 和 1.0 上给出 5 位有效数字的计算解. 把你的结果与精确解 $y=1/(1+x^2)$ 作比较.

556

4. 在你的计算机中心程序库中寻找一个适合求解一阶常微分方程系统初值问题的程序. 利用此程序求解下列问题:

$$\begin{cases} dx/dt = \sin x + \cos(yt) & x(-1) = 2.37 \\ dy/dt = e^{-xt} + [\sin(yt)]/t & y(-1) = -3.48 \end{cases}$$

在区间[−1, 4]上的阶要求具有 8 个小数位的精度. 在 t 轴上以 0.1 的间隔打印解.

a. 对 $-1 \leqslant t \leqslant 4$, 得到函数 $x(t)$ 和 $y(t)$ 的图表. 两条曲线应该出现在同一个图表上. 最大和最小坐标应该取 +4 和 −4.

b. 得到下列参数定义的曲线图

$$\{(x(t), y(t)): -1 \leqslant t \leqslant 4\}$$

(t 的值不出现在图中.) 这里得到的曲线称为原微分方程组的**轨道**.

8.5 局部误差和整体误差: 稳定性

8.4 节的亚当斯-巴什福思和亚当斯-莫尔顿公式仅仅是求解初值问题

$$\begin{cases} x' = f(t,x) \\ x(t_0) = x_0 \end{cases} \tag{1}$$

的多步法的两个例子. 术语**多步**是指在 x_n 的计算中需要某些前面算出的项 x_{n-1}, x_{n-2}, ….我们回顾上节中讨论过的一些内容.

8.5.1 隐式/显式以及收敛方法

任何多步法都可用下列形式的等式描述

$$a_k x_n + a_{k-1} x_{n-1} + \cdots + a_0 x_{n-k} = h[b_k f_n + b_{k-1} f_{n-1} + \cdots + b_0 f_{n-k}] \tag{2}$$

例如, 五阶亚当斯-莫尔顿公式为

$$x_n - x_{n-1} = h\left[\frac{251}{720} f_n + \frac{646}{720} f_{n-1} - \frac{264}{720} f_{n-2} + \frac{106}{720} f_{n-3} - \frac{19}{720} f_{n-4}\right] \tag{3}$$

在这些式子中, f_i 表示 $f(t_i, x_i)$. 在利用像(2)式那样的公式求解问题(1)时, 假定通过其他方法已经求得初值 x_0, x_1, x_2, …, x_{k-1}. 然后, 依次取 $n=k$, $k+1$, …使用(2)式. 我们看到当 $b_k \neq 0$ 时, (2)式两边同时包含未知坐标 x_n. 此时, 方法称为**隐式的**. 若 $b_k=0$, 则方法是**显式的**. 在我们的分析中, 假定 x_n 满足(2)式. 特别地, x_n 可从一个预估公式给出的尝试值出发, 通过迭代得到.

557

与(2)式对应的两个多项式是

$$\begin{cases} p(z) = a_k z^k + a_{k-1} z^{k-1} + \cdots + a_0 \\ q(z) = b_k z^k + b_{k-1} z^{k-1} + \cdots + b_0 \end{cases} \tag{4}$$

下面的分析将指出多步法的一些合乎需要的性质与多项式 p 和 q 的根的位置有关.

由(2)式定义的多步法在下面的情况之下称为**收敛的**：设想问题(1)的数值解是由变步长计算的. 我们用 $x(h, t)$ 表示用步长 h 得到的近似解. 通常，精确解写成 $x(t)$. 收敛指的是对某个区间 $[t_0, t_m]$ 中的一切 t，

$$\lim_{h\to 0} x(h,t) = x(t) \qquad (t\text{ 固定}) \tag{5}$$

只规定初值服从相同的等式，即，

$$\lim_{h\to 0} x(h, t_0 + nh) = x_0 \qquad (0 \leqslant n < k) \tag{6}$$

并且函数 f 满足基本的存在性定理的假设(8.1 节定理 3).

8.5.2 稳定性和相容性

所用的其他两个术语是稳定和相容. 若 p 的一切根位于圆盘 $|z| \leqslant 1$ 中且模为 1 的根是单根，则方法是**稳定的**. 若 $p(1)=0$ 且 $p'(1)=q(1)$，则方法是**相容的**. 现在可叙述这个主题中的主要定理.

定理 1(多步法稳定性和相容性定理) *(2)式的多步法收敛的充分必要条件是这个多步法是稳定的和相容的.*

证明 (稳定性是必要的.) 假如方法不稳定. 则不是 p 有一个根 λ 满足 $|\lambda| > 1$ 就是 p 有一个根 λ 满足 $|\lambda| = 1$ 且 $p'(\lambda)=0$. 在任一种情况，我们考虑其解是 $x(t)=0$ 的一个简单的初值问题：

$$\begin{cases} x' = 0 \\ x(0) = 0 \end{cases} \tag{7}$$

多步法是由等式

[558]

$$a_k x_n + a_{k-1} x_{n-1} + \cdots + a_0 x_{n-k} = 0 \tag{8}$$

所决定的. 这是一个线性差分方程，它的一个解是 $x_n = h\lambda^n$，其中 λ 是 p 的一个根. 若 $|\lambda| > 1$，则对 $0 \leqslant n < k$，我们有

$$|x(h, nh)| = h|\lambda^n| < h|\lambda|^k \to 0 \quad (\text{当 } h \to 0)$$

这建立了条件(6). 但违背(5)式，因为若 $t = nh$，则 $h = tn^{-1}$ 且

$$|x(h,t)| = |x(h,nh)| = tn^{-1}|\lambda|^n \to \infty$$

另一方面，若 $|\lambda|=1$ 且 $p'(\lambda)=0$，则(8)式的一个解是 $x_n = hn\lambda^n$. 条件(6)是满足的，因为当 $0 \leqslant n < k$ 时，

$$|x(h,nh)| = hn|\lambda|^n = hn < hk \to 0 \quad (\text{当 } h \to 0)$$

这是违背条件(5)的，因为

$$|x(h,t)| = (tn^{-1})n|\lambda|^n = t \neq 0$$

■

证明 (相容性是必要的.)假如(2)式定义的方法是收敛的. 考虑问题

$$\begin{cases} x' = 0 \\ x(0) = 1 \end{cases} \tag{9}$$

精确解是 $x=1$. (2)式再次具有形式(8). 取 $x_0 = x_1 = \cdots = x_{k-1} = 1$ 得到一个解，然后利用(8)式生成其余值 x_k，x_{k+1}，…. 因为方法收敛，所以 $\lim_{n\to\infty} x_n = 1$. 把它代入到(8)式得到结果 $a_k + a_{k-1} + \cdots + a_0 = 0$，或者换言之，$p(1)=0$.

现在考虑初值问题

$$\begin{cases} x' = 1 \\ x(0) = 0 \end{cases} \tag{10}$$

其精确解 $x=t$. (8)式变成

$$a_k x_n + a_{k-1} x_{n-1} + \cdots + a_0 x_{n-k} = h[b_k + b_{k-1} + \cdots + b_0] \tag{11}$$

因为方法收敛，因此根据前面的证明它是稳定的. 因而 $p(1)=0$ 且 $p'(1)\neq 0$. 由 $x_n=(n+k)h\gamma$, $\gamma=q(1)/p'(1)$给出(11)式的一个解. 实际上，把这个解代入到(11)式的左边得到

$$\begin{aligned} & h\gamma[a_k(n+k) + a_{k-1}(n+k-1) + \cdots + a_0 n] \\ &= nh\gamma(a_k + a_{k-1} + \cdots + a_0) + h\gamma[ka_k + (k-1)a_{k-1} + \cdots + a_1] \\ &= nh\gamma p(1) + h\gamma p'(1) \\ &= h\gamma p'(1) = hq(1) = h[b_k + b_{k-1} + \cdots + b_0] \end{aligned}$$

注意，因为$\lim\limits_{h\to 0}(n+k)h\gamma=0$，$n=0, 1, \cdots, k-1$，所以在这个数值解中的开始值与初值 $x(0)=0$ [559] 相容. 现在的收敛条件要求当 $nh=t$ 时，$\lim\limits_{n\to\infty} x_n=t$. 因此，我们有 $\lim\limits_{n\to\infty}(n+k)h\gamma=t$. 因为 $\lim\limits_{n\to\infty} kh=0$,所以我们得到 $\gamma=1$，或者 $p'(1)=q(1)$. ■

稳定性和相容性一起推出收敛性的证明是非常非常复杂的. 有兴趣的读者可查阅 Henrici [1962，第 5.3 节].

8.5.3 米尔恩方法

为说明多步法稳定性和相容性的定理 1，我们分析由

$$x_n - x_{n-2} = h\left[\frac{1}{3}f_n + \frac{4}{3}f_{n-1} + \frac{1}{3}f_{n-2}\right] \tag{12}$$

定义的**米尔恩方法**. 这是一个隐式方法，它由两个多项式

$$p(z) = z^2 - 1$$

$$q(z) = \frac{1}{3}z^2 + \frac{4}{3}z + \frac{1}{3}$$

来描述. 注意，p 的根是$+1$ 和-1. 它们都是单根. 而且 $p'(z)=2z$，$p'(1)=2$，$q(1)=2$. 因此，相容性和稳定性条件满足. 由定理可知米尔恩方法收敛.

8.5.4 局部截断误差

我们下面的工作是定义和分析在利用多步法(2)中产生的局部截断误差. 假定在所有前面的值 x_{n-1}，x_{n-2}，…是正确的假设下，利用(2)式来计算 x_n；即 $x_i=x(t_i)$，$i<n$. 这里 $x(t)$表示微分方程的解. 然后定义**局部截断误差**为 $x(t_n)-x_n$. 该误差完全是由差分方程模拟微分方程产生的. 在这个定义中，不包括舍入误差；假定 x_n 是从差分方程(2)用全精度算出的. 我们要证明：若方法有 m 阶(如 8.4 节中定义的)，则局部截断误差将是 $\mathcal{O}(h^{m+1})$. 因为我们的分析预先假定 f 和精确解函数 $x(t)$都具有一定的光滑性，所以在无限制的情况下，这个关于局部截断误差的结论不成立.

定理 2(*多步法的局部截断误差定理*) *若多步法(2)是 m 阶的，$x\in C^{m+2}$ 且 $\partial f/\partial x$ 连续，则在前段的假设之下，*

$$x(t_n) - x_n = \left(\frac{d_{m+1}}{a_k}\right)h^{m+1}x^{(m+1)}(t_{n-k}) + \mathcal{O}(h^{m+2}) \tag{13}$$

560 (系数 d_k 是8.4节中定义的.)

证明 只要证明 $n=k$ 时的等式就足够了，因为 x_n 可解释为在点 t_{n-k} 上开始的数值解的值. 利用8.4节中(10)式的线性泛函 L，我们有

$$Lx = \sum_{i=0}^{k}[a_i x(t_i) - hb_i x'(t_i)] = \sum_{i=0}^{k}[a_i x(t_i) - hb_i f(t_i, x(t_i))] \tag{14}$$

另一方面，数值解满足等式

$$0 = \sum_{i=0}^{k}[a_i x_i - hb_i f(t_i, x_i)] \tag{15}$$

因为我们已经假定 $x_i = x(t_i)$，$i<k$，从(14)式减去(15)式的结果是

$$Lx = a_k[x(t_k) - x_k] - hb_k[f(t_k, x(t_k)) - f(t_k, x_k)] \tag{16}$$

对(16)式的最后一项应用中值定理，得到

$$\begin{aligned} Lx &= a_k[x(t_k) - x_k] - hb_k \frac{\partial f}{\partial x}(t_k, \xi)[x(t_k) - x_k] \\ &= [a_k - hb_k F][x(t_k) - x_k] \end{aligned} \tag{17}$$

其中对 $x(t_k)$ 和 x_k 之间的某个 ξ，$F = \partial f(t_k, \xi)/\partial x$. 回顾8.4节中的(10)式，若使用的方法为 m 阶，则 Lx 将有形式

$$Lx = d_{m+1}h^{m+1}x^{(m+1)}(t_0) + \mathcal{O}(h^{m+2}) \tag{18}$$

组合(17)式和(18)式，得到(13)式. 这里可略去分母中的 hb_kF. (见习题8.5.8.) ■

8.5.5 整体截断误差

现在我们希望建立微分方程数值求解中的整体截断误差界. 在求解过程中任何给定的 t_n 步上，用 x_n 表示计算解. 假定所有的计算是用全精度执行的；即不包含舍入误差. 当然，在 t_n 的真解 $x(t_n)$ 不同于计算解 x_n，因为后者是用近似泰勒级数的公式得到的. 差 $x(t_n)-x_n$ 是**整体截断误差**. 它不单单是前面的点上出现的所有局部截断误差之和. 为看清这一点，我们必须了解数值解的每一步必须使用前面步上计算的近似纵坐标作为它的初值. 因为纵坐标是带误差的，所以数值过程实际上是试图跟踪错误的解曲线. 因此我们必须着手分析，看看如果从不561 同的初始条件出发两条解曲线有怎样的差别. 换言之，我们需要了解改变解曲线纵坐标之后的初值的影响. 图8-2说明我们正试图比较的内容.

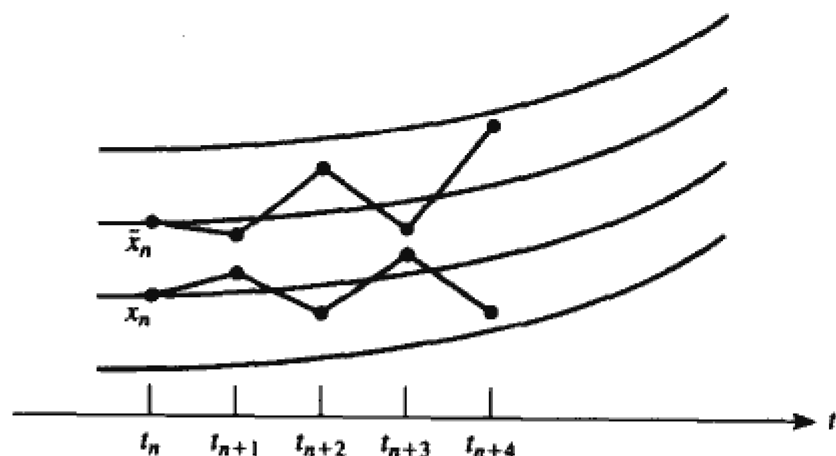


图8-2 改变初始条件的影响

考虑初值问题

$$\begin{cases} x' = f(t,x) \\ x(0) = s \end{cases} \tag{19}$$

这里用 f_x 表示 $\partial f / \partial x$，假定它是连续的并在 $0 \leqslant t \leqslant T$ 和 $-\infty < x < +\infty$ 定义的区域内满足条件 $f_x(t, x) \leqslant \lambda$. (19)的解是 t 的函数，但为了表明它与初值 s 有关，我们记它为 $x(t; s)$. 定义 $u(t) = \partial x(t; s) / \partial s$. 对初值问题(19)中 u 关于 s 的微分可得到一个微分方程——**变分方程**. 其结果是

$$\begin{cases} u'(t) = f_x(t,x)u \\ u(0) = 1 \end{cases} \tag{20}$$

例 1 在初值问题

$$\begin{cases} x' = x^2 \\ x(0) = s \end{cases} \tag{21}$$

中显式地求出 u.

解 这里 $f(t, x) = x^2$，$f_x = 2x$. 因此，变分方程是

$$\begin{cases} u' = 2xu \\ u(0) = 1 \end{cases} \tag{22}$$

初值问题(21)的解是 $x(t) = s(1-st)^{-1}$. 因此(22)式变成

$$\begin{cases} u'(t) = 2s(1-st)^{-1}u(t) \\ u(0) = 1 \end{cases}$$

这个问题的解是

$$u(t) = (1-st)^{-2}$$

■ 562

定理 3(变分方程定理) *若 $f_x \leqslant \lambda$，则变分方程(20)的解满足不等式*

$$|u(t)| \leqslant e^{\lambda t} \qquad (t \geqslant 0)$$

证明 从(20)式，我们有

$$u'/u = f_x = \lambda - \alpha(t) \tag{23}$$

其中 $\alpha(t) \geqslant 0$. 对(23)式进行积分，得到

$$\log|u| = \lambda t - \int_0^t \alpha(\tau)\mathrm{d}\tau = \lambda t - A(t)$$

其中 $A(t)$ 表示显示的积分. 因为 $t \geqslant 0$，所以 $A(t) \geqslant 0$，因此 $\log|u| \leqslant \lambda t$. 因为指数函数是递增的，所以 $|u| \leqslant e^{\lambda t}$. ■

定理 4(初值问题解曲线定理) *若初值问题(19)用初值 s 和 $s+\delta$ 求解，则解曲线在 t 的差别至多为 $|\delta| e^{\lambda t}$.*

证明 对变分方程用中值定理、u 的定义和定理 3，我们有

$$|x(t;s) - x(t;s+\delta)| = \left|\frac{\partial}{\partial s}x(t,s+\theta\delta)\right||\delta|$$

$$= |u(t)||\delta| \leqslant |\delta| e^{\lambda t}$$

■

定理 5(整体截断误差界定理) *若在 t_1，t_2，…，t_n 上的局部截断误差在数量上不超过 δ，*

则在 t_n 上的整体截断误差不超过 $\delta(e^{n\lambda h}-1)(e^{\lambda h}-1)^{-1}$.

证明 设 δ_1，δ_2，…是对应于点 t_1，t_2，…上的数值解的截断误差. 在计算 x_2 的过程中，初始条件有一个 δ_1 的误差，而由定理 4，在初值问题的解曲线上这个误差在 t_2 的影响至多为 $|\delta_1|e^{\lambda h}$. 把这个值加到 t_2 的截断误差上. 因此，t_2 上的整体截断误差至多是

$$|\delta_1|e^{\lambda h}+|\delta_2|$$

这个误差在 t_3 上的影响(由定理 4)不大于

$$(|\delta_1|e^{\lambda h}+|\delta_2|)e^{\lambda h}$$

把这个值加到 t_3 的截断误差上. 以这个方式继续下去，我们求出在 t_n 上的整体截断误差不大于

563

$$\sum_{k=1}^{n}|\delta_k|e^{(n-k)\lambda h}\leqslant\delta\sum_{k=0}^{n-1}e^{k\lambda h}=\delta(e^{n\lambda h}-1)(e^{\lambda h}-1)^{-1}$$ ■

定理 6(整体截断误差逼近定理) *若数值解中局部截断误差是 $\mathcal{O}(h^{m+1})$，则整体截断误差是 $\mathcal{O}(h^m)$.*

证明 在整体截断误差界的定理 5 中，设 δ 是 $\mathcal{O}(h^{m+1})$. 因为 e^z-1 是 $\mathcal{O}(z)$且 $nh=t$，利用定理 5 中的公式我们发现阶减少 1. ■

习题 8.5

1. 按照多步法稳定性和相容性的定理 1. 讨论下列这些多步法：

   a. $x_n-x_{n-2}=2hf_{n-1}$

   b. $x_n-x_{n-2}=h[\frac{7}{3}f_{n-1}-\frac{2}{3}f_{n-2}+\frac{1}{3}f_{n-3}]$

   c. $x_n-x_{n-1}=h[\frac{3}{8}f_n+\frac{19}{24}f_{n-1}-\frac{5}{24}f_{n-2}+\frac{1}{24}f_{n-3}]$

2. 一个方法称为**弱稳定的**，如果 p 有一个零点 λ 使得 $\lambda\neq1$，$|\lambda|=1$ 且 $q(\lambda)<\lambda p'(\lambda)$. 证明(12)式给出的米尔恩方法是**弱不稳定的**.

3. 证明：每一个多步法，其中 $p(z)=z^k-z^{k-1}$，$\sum_{i=0}^{k}b_i=1$，是稳定的，相容的，收敛的和弱稳定的.

4. 确定下列多步法的数值特征.

$$x_n+4x_{n-1}-5x_{n-2}=h[4f_{n-1}+2f_{n-2}]$$

5. 是否存在不信任求解 $x'=f(t,x)$的数值方法

$$x_n-3x_{n-1}+2x_{n-2}=h[f_n+2f_{n-1}+f_{n-2}-2f_{n-3}]$$

   的理由？为什么？

6. 下列多步法中哪一个是收敛的？

   a. $x_n-x_{n-2}=h(f_n-3f_{n-1}+4f_{n-2})$

   b. $x_n-2x_{n-1}+x_{n-2}=h(f_n-f_{n-1})$

   c. $x_n-x_{n-1}-x_{n-2}=h(f_n-f_{n-1})$

   d. $x_n-3x_{n-1}+2x_{n-2}=h(f_n+f_{n-1})$

   e. $x_n-x_{n-2}=h(f_n-3f_{n-1}+2f_{n-2})$

7. 多步法称为**强稳定的**，如果 $p(1)=0$，$p'(1)\neq0$ 且 p 的所有其他根满足不等式 $|z|<1$. 利用多步法稳定性和相容性定理 1，证明强稳定方法是收敛的. 再证明对任何 λ 的值，强稳定方法求解问题 $x'=\lambda x$，$x(0)=1$

没有引入任何额外的指数增长的误差.

8. 证明

$$\frac{Ah^{m+1}+\mathcal{O}(h^{m+2})}{B-Ch}=\frac{A}{B}h^{m+1}+\mathcal{O}(h^{m+2})$$

564

8.6 方程组和高阶常微分方程

一阶微分方程组的标准形式为

$$\begin{cases}x_1'=f_1(t,x_1,x_2\cdots,x_n)\\x_2'=f_2(t,x_1,x_2,\cdots,x_n)\\\qquad\vdots\\x_n'=f_n(t,x_1,x_2,\cdots,x_n)\end{cases}\tag{1}$$

在这个方程组中，要确定 n 个未知函数 x_1，x_2，…，x_n. 它们是单变量 t 的函数，记号 x_i' 表示导数 dx_i/dt. 考虑这个方程组的一个具体例子，其中用 x 和 y 代替 x_1 和 x_2：

$$\begin{cases}x'=x+4y-e^t\\y'=x+y+2e^t\end{cases}\tag{2}$$

方程组(2)的通解是

$$x=2ae^{3t}-2be^{-t}-2e^t$$

$$y=ae^{3t}+be^{-t}+\frac{1}{4}e^t\tag{3}$$

其中 a 和 b 是任意常数. 当然，通过微分并代入方程组(2)可以验证假定的解. 注意，此例是未知函数 x 和 y 的线性方程组.

在估计可能具有唯一解的明确定义的物理问题中，微分方程组会伴有确定通解中任意常数的辅助条件. 因此，若方程组(2)伴随初始条件

$$x(0)=4,y(0)=\frac{5}{4}$$

则解为

$$\begin{cases}x=4e^{3t}+2e^{-t}-2e^t\\y=2e^{3t}-e^{-t}+\frac{1}{4}e^t\end{cases}\tag{4}$$

一般的方程组(1)的初值问题由 n 个微分方程连同给定的初值 $t=t_0$ 以及详细说明每个函数 x_i 在 t_0 的值所组成.

8.6.1 向量记号

一种方便的向量记号可用来改写方程组(1). 设 X 表示其分量为 x_1，x_2，…，x_n 的列向量. 这些分量是 t 的函数. 因此 X 是$\mathbb{R}$(或$\mathbb{R}$中的一个区间)到$\mathbb{R}^n$的一个映射. 类似地，设 F 表示具有分量 f_1，f_2，…，f_n 的列向量. 每个分量是$\mathbb{R}^{n+1}$(或它的一个子集)上的一个函数，故 F 是$\mathbb{R}^{n+1}$到$\mathbb{R}^n$的一个映射. 方程组(1)可写成

565

$$X'=F(t,X)\tag{5}$$

方程组(5)的初值问题还包括向量 $X(t_0)$的数值，其中 t_0 是 t 的初值.

高阶微分方程可以转换为一阶微分方程组. 假定给出下列形式的单个微分方程

$$y^{(n)} = f(t, y, y', y'', \cdots, y^{(n-1)})$$

当然，这里所有的导数是关于 t 的：$y^{(i)} = \mathrm{d}^i y / \mathrm{d} t^i$. 我们按照下面这些定义

$$x_1 = y \quad x_2 = y' \quad x_3 = y'' \quad \cdots \quad x_n = y^{(n-1)}$$

引进新变量 x_1，x_2，…，x_n. 新变量满足一阶微分方程组：

$$\begin{cases} x'_1 = x_2 \\ x'_2 = x_3 \\ x'_3 = x_4 \\ \quad \vdots \\ x'_n = f(t, x_1, x_2, x_3, \cdots, x_n) \end{cases}$$

这是一个(5)式所示形式的方程组.

为利用广泛的可得到的软件求解微分方程，把问题转换成像(5)式那样的方程组几乎总是必要的. 我们用两个例子来说明这个过程.

例 1 把初值问题

$$\begin{cases} (\sin t) y''' + \cos(ty) + \sin(t^2 + y'') + (y')^3 = \log t \\ y(2) = 7 \\ y'(2) = 3 \\ y''(2) = -4 \end{cases} \tag{6}$$

转换为有初值的一阶微分方程组.

解 我们引进新变量 x_1，x_2 和 x_3 如下：$x_1 = y$，$x_2 = y'$，$x_3 = y''$. 控制 $X = (x_1,\ x_2,\ x_3)^{\mathrm{T}}$ 的方程组是

$$\begin{cases} x'_1 = x_2 \\ x'_2 = x_3 \\ x'_3 = [\log t - x_2^3 - \sin(t^2 + x_3) - \cos(t x_1)] / \sin t \end{cases} \tag{7}$$

566 在 $t=2$ 的初始条件是 $X = (7,\ 3,\ -4)^{\mathrm{T}}$. ■

高阶方程组可以用下例中说明的同样的方式处理.

例 2 把下列方程组

$$\begin{cases} (x'')^2 + t\mathrm{e}^y + y' = x' - x \\ y'y'' - \cos(xy) + \sin(tx'y) = x \end{cases} \tag{8}$$

转换成一阶微分方程组. 初始条件在此例中被略去.

解 在引进新变量 $x_1 = x$，$x_2 = x'$，$x_3 = y$ 和 $x_4 = y'$之后，问题可写成

$$\begin{cases} x'_1 = x_2 \\ x'_2 = (x_2 - x_1 - x_4 - t\mathrm{e}^{x_3})^{1/2} \\ x'_3 = x_4 \\ x'_4 = [x_1 - \sin(t x_2 x_3) + \cos(x_1 x_3)] / x_4 \end{cases} \tag{9}$$

■

8.6.2 方程组的泰勒级数方法

8.2 节中讨论过的泰勒级数方法可应用于一阶方程组. 对每个变量我们写出如下的截断泰勒级数：

$$x_i(t+h) = x_i(t) + hx'_i(t) + \frac{h^2}{2!}x''_i(t) + \frac{h^3}{3!}x'''_i(t) + \cdots + \frac{h^n}{n!}x_i^{(n)}(t)$$

或按向量记号：

$$X(t+h) = X(t) + hX'(t) + \frac{h^2}{2!}X''(t) + \frac{h^3}{3!}X'''(t) + \cdots + \frac{h^n}{n!}X^{(n)}(t) \tag{10}$$

这里出现的导数可以从微分方程得到. 通常当这些导数用于一个计算机程序中时必须以一个特别的次序计算它们. 我们必须保证在一步中需要的量作为前面步的结果是可利用的.

例 3 对下列初值问题编写三阶泰勒级数算法. 利用 $|h|=0.1$ 并在区间 $-2\leqslant t\leqslant 1$ 上计算解.

$$\begin{cases} x' = x + y^2 - t^3 & x(1) = 3 \\ y' = y + x^3 + \cos t & y(1) = 1 \end{cases} \tag{11}$$

解 需要的高阶导数是

$$\begin{aligned} x'' &= x' + 2yy' - 3t^2 \\ y'' &= y' + 3x^2x' - \sin t \\ x''' &= x'' + 2yy'' + 2(y')^2 - 6t \\ y''' &= y'' + 6x(x')^2 + 3x^2x'' - \cos t \end{aligned}$$

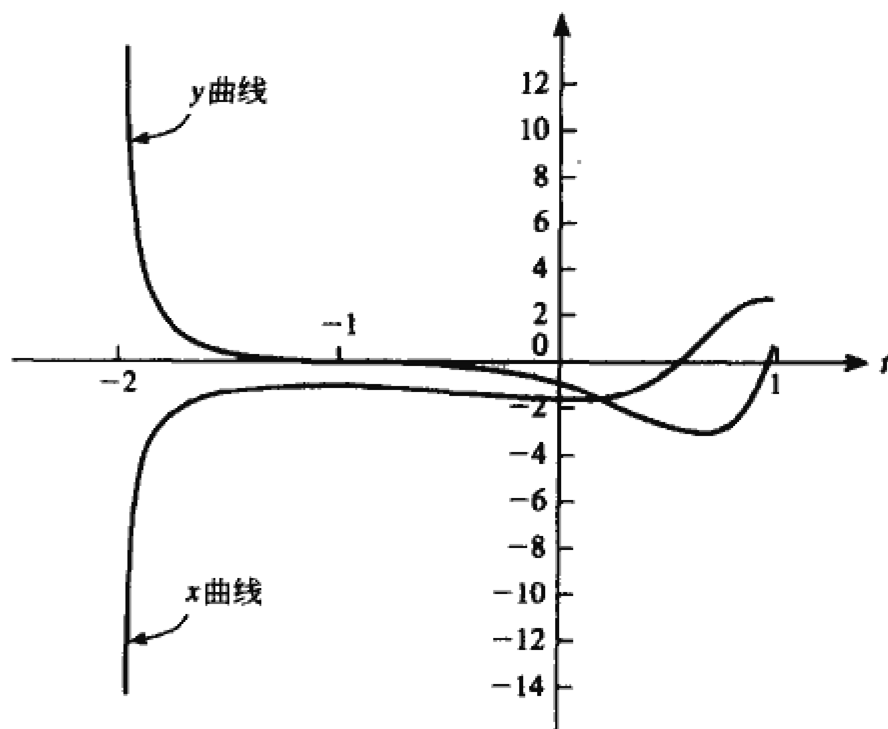
567

执行计算的一个适当的算法如下：

```
input t←1; x←3; y←1; h←−0.1; M←30
output 0, t, x, y
for k=1 to M do
    x'←x+y²−t³
    y'←y+x³+cost
    x''←x'+2yy'−3t²
    y''←y'+3x²x'−sint
    x'''←x''+2yy''+2(y')²−6t
    y'''←y''+6x(x')²+3x²x''−cost
    x←x+h(x'+(1/2)h(x''+(1/3)h(x''')))
    y←y+h(y'+(1/2)h(y''+(1/3)h(y''')))
    t←t+h
    output k, t, x, y
end do
```

在这里不给出计算机实施这个算法的数值输出，但是在图 8-3 中给出两个函数 $x(t)$ 和 $y(t)$ 的计算机图形. 因为绘图程序一点也不标准，所以给出得到图形的计算机程序毫无意义. 通常我们必须给出自动绘图机两个数组 $\{t_i: 0\leqslant i\leqslant m\}$ 和 $\{x_i: 0\leqslant i\leqslant m\}$，然后由此在笛卡儿网格点上绘出点. 通常可以命令绘图机用直线连接这些点. 绝对要注意点的排序. 当然，通常为

$t_0 < t_1 < \cdots < t_m$. 如果绘图点互相靠近的话，则所得的曲线好像是一条光滑曲线而不是一串线段序列.



568

图 8-3 例 3 的解曲线

理论上，不失一般性，假设方程组(1)中的方程不显式地包含 t. 我们引入一个变量 $x_0 = t$ 可把方程组写成下列形式

$$x'_i = f_i(x_0, x_1, \cdots, x_n)$$

关于新变量的微分方程就是 $x'_0 = 1$. 这样，我们可把方程组(5)写成

$$X' = F(X) \tag{12}$$

t 不显式地出现，这里 $X = (x_0, x_1, \cdots, x_n)^T$. 形式为(12)的方程组称为**自控的**.

例 4 把例 1 中的初值问题转换成 t 不显式出现的方程组.

解 设 x_1，x_2 和 x_3 如例 1 中那样. 取 $x_0 = t$. 新方程组是

$$\begin{cases} x'_0 = 1 \\ x'_1 = x_2 \\ x'_2 = x_3 \\ x'_3 = [\log x_0 - x_2^3 - \sin(x_0^2 + x_3) - \cos(x_0 x_1)]/\sin x_0 \end{cases} \tag{13}$$

而初始条件是 $X = (2, 7, 3, -4)^T$.

8.6.3 方程组的其他方法

当方程组是自控的时候，即它具有(12)式的形式时，最容易写出一阶方程组的龙格-库塔方法. 经典的向量形式的**四阶龙格-库塔公式**是

$$X(t+h) = X(t) + \frac{1}{6}(F_1 + 2F_2 + 2F_3 + F_4) \tag{14}$$

其中

$$\begin{cases} F_1 = hF(X) \\ F_2 = hF(X + \frac{1}{2}F_1) \\ F_3 = hF(X + \frac{1}{2}F_2) \\ F_4 = hF(X + F_3) \end{cases}$$

执行这个过程的子程序的程序设计留作计算机习题 8.6.4. 对 8.3 节中讨论的龙格-库塔-费尔贝格方法可给出类似的公式.

多步法也可推广到应用于方程组. 例如，我们给出 8.4 节中**亚当斯-巴什福思-莫尔顿预估-校正方法**(4)和(8)的向量形式：

569

$$X^*(t+h) = X(t) + \frac{h}{720}[1\,901F(X(t)) - 2\,774F(X(t-h)) + 2\,616F(X(t-2h)) - 1\,274F(X(t-3h)) + 251F(X(t-4h))]$$

$$X(t+h) = X(t) + \frac{h}{720}[251F(X^*(t+h)) + 646F(X(t)) - 264F(X(t-h)) + 106F(X(t-2h)) - 19F(X(t-3h))]$$

正如在单个方程情况中那样，一个单步过程，譬如可用五阶龙格-库塔方法(计算机习题 8.3.7～8.3.9)提供起始值：

$$X(t_0+h) \qquad X(t_0+2h) \qquad X(t_0+3h) \qquad X(t_0+4h)$$

习题 8.6

1. 求下列方程组的通解：

$$\begin{cases} x' = 3x - 4y + e^t \\ y' = x - y - e^t \end{cases}$$

提示：尝试形式为 e^t，te^t 和 t^2e^t 的函数.

2. 写出等价于

$$\begin{cases} x''' - \sin(x'') + e^t x' + 2t\cos x = 25 \\ x(0) = 5 \qquad x'(0) = 3 \qquad x''(0) = 7 \end{cases}$$

的一个自控的一阶方程组.

3. 把三阶常微分方程

$$\begin{cases} x''' + 2x'' - x' - 2x = e^t \\ x(8) = 3 \qquad x'(8) = 2 \qquad x''(8) = 1 \end{cases}$$

写成一个自控的一阶方程组.

4. 把二阶常微分方程组

$$\begin{cases} x'' - x'y = 3y'x\log t \\ y'' - 2xy' = 5x'y\sin t \end{cases}$$

转换成 t 不显式出现的一阶方程组.

5. 用向量记号把

$$\begin{cases} y'' + yz = 0 \qquad y(0) = 1 \qquad y'(0) = 0 \\ z' + 2yz = 4 \qquad z(0) = 3 \end{cases}$$

写成一个带初始条件的一阶方程组.

6. 写出等价于

$$\begin{cases} x''' - [\sin x'' + e^t x']^2 + \cos x = 0 \\ x(0) = 3 \qquad x'(0) = 4 \qquad x''(0) = 5 \end{cases}$$

[570] 的一个自控的一阶方程组.

7. 假定可利用求解一阶方程组初值问题的计算机程序. 说明如何数值求解下列问题:

$$\begin{cases} x'' = x\cos t + e^t x' + 3t^2 + 7 \\ x(1) = 5 \qquad x'(1) = 9 \end{cases}$$

8. 把下列高阶常微分方程组

$$\begin{cases} y_1^{(n)} = f_1(t, y_1, y_1', \cdots, y_1^{(n-1)}) \\ y_2^{(n)} = f_2(t, y_2, y_2', \cdots, y_2^{(n-1)}) \\ \qquad \vdots \\ y_m^{(n)} = f_m(t, y_m, y_m', \cdots, y_m^{(n-1)}) \end{cases}$$

写成一个一阶方程组.

9. 把下列高阶微分方程组

$$\begin{cases} x''' - 5tx''y'' + \ln(x')z = 0 \\ y'' - \sin(ty) + 7tx'' = 0 \\ z' + 16ty' - e^t zx' = 0 \end{cases}$$

转换成一个 t 不显式出现的一阶方程组.

10. 写出(12)式形式的方程组的欧拉方法、Heun 方法和修正的欧拉方法.

计算机习题 8.6

1. 利用泰勒级数方法写出求解初值问题

$$\begin{cases} x_1' = t + x_1^2 + x_2 \qquad x_1(-1) = 0.43 \\ x_2' = t^2 - x_1 + x_2^2 \qquad x_2(-1) = -0.69 \end{cases}$$

的计算机程序. 要包括 h, h^2 和 h^3 项, 并连续解到 $t=1$. 设 $h=0.01$.

2. (挑战性习题)对初值问题

$$\begin{cases} x_1' = \sin x_1 + \cos(tx_2) \qquad x_1(-1) = 2.37 \\ x_2' = t^{-1}\sin(tx_1) \qquad x_2(-1) = -3.48 \end{cases}$$

执行上题中的指令. 因为在 $t=0$ 上的奇异性, 所以 x_2', x_2'' 和 x_2''' 的程序设计必须仔细地进行.

3. 数值求解微分方程组

$$\begin{cases} x_1' = -13x_1 + 6x_2 \qquad x_1(0) = -12 \\ x_2' = -13x_2 - 6x_1 \qquad x_2(0) = 6 \end{cases}$$

其中 $c=\cos 6t$, $s=\sin 6t$. 在区间[0, 10]上用步长 $h=0.01$ 积分. 验证正确解是

$$\begin{cases} x_1 = e^{-13t}(s - 2c) \\ x_2 = e^{-13t}(2s + c) \end{cases}$$

[571] 你的计算解与真解比较满意吗? 这个例子是由 Lambert[1973]提供的.

4. 编写一个步长为 h 的四阶龙格-库塔方法的子程序或过程, 使它能处理 $n \leqslant 20$ 的一个 n 维微分方程组. 利用 $h=-0.01$, 通过在区间 $1 \leqslant t \leqslant 2$ 上求解方程组

$$\begin{cases} x' = x^{-2} + \log y + t^2 \\ y' = e^y - \cos x + (\sin t)x - (xy)^{-3} \\ x(2) = -2 \qquad y(2) = 1 \end{cases}$$

测试你的程序.

5. 对常微分方程 $x''+192x=0$，$x(0)=\frac{1}{6}$，$x'(0)=0$ 求解并在区间[0，5]上画出结果曲线. 这对应于一个**无阻尼弹性-质量系统**.
6. 编写且测试能处理一阶微分方程组的自适应龙格-库塔-费尔贝格算法的子程序或过程.
7. 编写且测试能处理一阶微分方程组的结合五阶龙格-库塔方法(见计算机习题 8.3.7)的五阶亚当斯-巴什福思-莫尔顿方法的子程序或过程.

8.7 边值问题

在本章的前面几节中，对求解初值问题考察了各种方法，例如，初值问题

$$\begin{cases} x'' = f(t,x,x') \\ x(a) = \alpha \qquad x'(a) = \beta \end{cases} \tag{1}$$

取 $x_1=x$ 和 $x_2=x'$时可以转换成一阶方程组的形式

$$\begin{cases} x_1' = x_2 & x_1(a) = \alpha \\ x_2' = f(t,x_1,x_2) & x_2(a) = \beta \end{cases} \tag{2}$$

然后，方程组(2)可用前面描述的步进方法之一求解.

然而，如果(1)的问题改变为

$$\begin{cases} x'' = f(t,x,x') \\ x(a) = \alpha \qquad x(b) = \beta \end{cases} \tag{3}$$

则情况完全不同. 差别在于在两个不同的点 $t=a$ 和 $t=b$ 上指定条件. 初值问题的步进方法不适合对(3)求解，因为没有完整的补充初值，数值求解不能开始. (3)式是**两点边值问题**的一个典型例子. 通常此类问题出现的困难大于初值问题的困难.

下面是一个可以不用数值方法求解的两点边值的例子. 572

$$\begin{cases} x'' = -x \\ x(0) = 3 \qquad x(\frac{\pi}{2}) = 7 \end{cases} \tag{4}$$

首先我们可求微分方程的通解，它是

$$x(t) = A\sin t + B\cos t \tag{5}$$

然后，可确定常数 A 和 B 使其满足边界条件. 因此

$$\begin{cases} 3 = x(0) = A\sin 0 + B\cos 0 = B \\ 7 = x(\frac{\pi}{2}) = A\sin\frac{\pi}{2} + B\cos\frac{\pi}{2} = A \end{cases}$$

于是，(4)的解是

$$x(t) = 7\sin t + 3\cos t$$

更多的此类例子在习题中给出.

如果(3)中的微分方程的通解不知道的话，则刚才说明的技巧便是无效的．这里，我们的兴趣是在于可着手处理任何两点边值问题的数值方法．

8.7.1 存在性

在考虑数值方法之前，我们适当地讨论一下存在性问题．一般说来，只假设 f 是一个好的函数并不能保证对(3)的解的存在性．支持这个论断的一个简单例子是

$$\begin{cases} x'' = -x \\ x(0) = 3 \qquad x(\pi) = 7 \end{cases} \tag{6}$$

这个问题与(4)中的问题只有一点点不同，但当试图把边界值加在通解(5)上时，我们得到矛盾的等式 $3=B$ 和 $7=-B$．因此(6)中的问题无解．

关于两点边值问题(3)解的存在性定理往往是相当复杂的，我们建议读者查阅 Stoer and Bulirsch[1980]与 Keller[1968]的著作．在 Keller[1968，第 108 页]中一个漂亮的结果是下列定理．

定理 1(边值问题存在性定理) 当 $\partial f/\partial x$ 连续、非负且在不等式 $0\leqslant t\leqslant 1$，$-\infty<x<+\infty$ 定义的无限带内有界时，边值问题

$$\begin{cases} x'' = f(t,x) \\ x(0) = 0 \qquad x(1) = 0 \end{cases}$$

573 有唯一解．

例 1 证明下列两点边值问题有唯一解．

$$\begin{cases} x'' = (5x + \sin 3x)e^t \\ x(0) = x(1) = 0 \end{cases}$$

解 尝试利用定理 1．于是

$$\frac{\partial f}{\partial x} = (5 + 3\cos 3x)e^t$$

它在带 $0\leqslant t\leqslant 1$，$-\infty<x<+\infty$ 内是连续的．而且，它以 8e 为上界，因为 $3\cos 3x\geqslant -3$，所有它是非负的．因此满足定理 1 的假设． ■

8.7.2 变量替换

虽然边值问题存在性定理 1 应用于一种特殊的情况，但是简单的变量替换可简化更一般的问题为特殊情况．为此，我们从改变 t 区间人手．假定原问题是

$$\begin{cases} x'' = f(t,x) \\ x(a) = \alpha \qquad x(b) = \beta \end{cases} \tag{7}$$

其中 $x=x(t)$．这里命名一个变量替换 $t=a+(b-a)s$．注意 $s=0$ 对应于 $t=a$，$s=1$ 对应于 $t=b$．所以，我们定义 $y(s)=x(a+\lambda s)$，$\lambda=b-a$．于是 $y'(s)=\lambda x'(a+\lambda s)$，$y''(s)=\lambda^2 x''(a+\lambda s)$．同样地，$y(0)=x(a)=\alpha$，$y(1)=x(b)=\beta$．因此，若 x 是(7)的解，则 y 是边值问题

$$\begin{cases} y''(s) = \lambda^2 f(a+\lambda s, y(s)) \\ y(0) = \alpha \qquad y(1) = \beta \end{cases} \tag{8}$$

的解．反之，若 y 是(8)的解，则函数 $x(t)=y((t-a)/(b-a))$是(7)的解．

定理 2(两点边值问题第一定理) 考察下列两点边值问题:

1. $\begin{cases} x''=f(t,x) \\ x(a)=\alpha \qquad x(b)=\beta \end{cases}$

2. $\begin{cases} y''=g(t,y) \\ y(0)=\alpha \qquad y(1)=\beta \end{cases}$

其中

$$g(p,q)=(b-a)^2 f(a+(b-a)p,q)$$

[574]

若 y 是问题 2 的解,则函数

$$x(t)=y((t-a)/(b-a))$$

是问题 1 的解. 而且,若 x 是问题 1 的解,则

$$y(t)=x(a+(b-a)t)$$

是问题 2 的解.

证明 下面是一个简单的验证:

$$x(a)=y\left(\frac{a-a}{b-a}\right)=y(0)=\alpha$$

$$x(b)=y\left(\frac{b-a}{b-a}\right)=y(1)=\beta$$

$$x'(t)=y'\left(\frac{t-a}{b-a}\right)\frac{1}{b-a}$$

$$x''(t)=y''\left(\frac{t-a}{b-a}\right)\frac{1}{(b-a)^2}$$

$$=g\left(\frac{t-a}{b-a},y\left(\frac{t-a}{b-a}\right)\right)\frac{1}{(b-a)^2}$$

$$=(b-a)^2 f\left(a+(b-a)\frac{t-a}{b-a},y\left(\frac{t-a}{b-a}\right)\right)\frac{1}{(b-a)^2}=f(t,x(t))$$

■

例 2 说明下列两个两点边值问题等价:

$$\begin{cases} x''=\sin(tx)+x^2 \\ x(1)=3 \qquad x(4)=7 \end{cases}$$

$$\begin{cases} y''=16\{\sin[(4s+1)y]+y^2\} \\ y(0)=3 \qquad y(1)=7 \end{cases}$$

解 利用定理 2. ■

为了简化两点边值问题

$$\begin{cases} y''=g(t,y) \\ y(0)=\alpha \qquad y(1)=\beta \end{cases}$$

成为一个具有齐次边值的问题,我们仅仅从 y 中减去在 0 和 1 取值为 α 和 β 的线性函数. 下面是一个有关的定理.

定理 3(两点边值问题第二定理) 考察下列两点边值问题:

[575]

1. $\begin{cases} y''=g(t, y) \\ y(0)=\alpha \qquad y(1)=\beta \end{cases}$

2. $\begin{cases} z''=h(t, z) \\ z(0)=0 \qquad z(1)=0 \end{cases}$

其中

$$h(p,q) = g(p,q+\alpha+(\beta-\alpha)p)$$

若 z 是问题 2 的解，则函数

$$y(t) = z(t)+\alpha+(\beta-\alpha)t$$

是问题 1 的解. 而且，若 y 是问题 1 的解，则

$$z(t) = y(t)-[\alpha+(\beta-\alpha)t]$$

是问题 2 的解.

证明 再次直接验证：

$$\begin{aligned} y(0) &= z(0)+\alpha+(\beta-\alpha)0 = \alpha \\ y(1) &= z(1)+\alpha+(\beta-\alpha)1 = \beta \\ y''(t) &= z''(t) = h(t,z(t)) = g(t,z(t)+\alpha+(\beta-\alpha)t) \\ &= g(t,y(t)) \end{aligned}$$

■

例 3 说明下列问题有唯一解：

$$\begin{cases} x'' = [5x-10t+35+\sin(3x-6t+21)]e^t \\ x(0) = -7 \qquad x(1) = -5 \end{cases}$$

解 因为这个问题中边界值不是齐次的，所以不能直接应用定理 1. 为得到 0 边界值的等价问题，我们如定理 3 中那样改变相关的变量. 设

$$z(t) = x(t)-\ell(t) \quad 和 \quad \ell(t) = -7+2t$$

则

$$\begin{aligned} z''=x''&=[5x-10t+35+\sin(3x-6t+21)]e^t \\ &=\{5(z+\ell)-10t+35+\sin[3(z+\ell)-6t+21]\}e^t \\ &=\{5z+\sin 3z\}e^t \end{aligned}$$

新变量 z 的边界值是

$$\begin{aligned} z(0) &= x(0)-\ell(0) = -7+7 = 0 \\ z(1) &= x(1)-\ell(1) = -5-(-5) = 0 \end{aligned}$$

576 正如例 1 中指出的那样，关于 z 的边值问题有唯一解. 因此，关于 x 的问题有唯一解. ■

例 4 把下列两点边值问题转换成在区间[0, 1]上有 0 边界值的等价问题.

$$\begin{cases} x'' = x^2+3-t^2+xt \\ x(3) = 7 \qquad x(5) = 9 \end{cases} \tag{9}$$

解 由定理 2，等价问题是

$$\begin{cases} y'' = g(t,y) \\ y(0) = 7 \qquad y(1) = 9 \end{cases} \tag{10}$$

其中

$$g(t,y)=4f(3+2t,y)=4[y^2+3-(3+2t)^2+y(3+2t)]$$

由定理 3，另一个等价问题是

$$\begin{cases}z''=h(t,z)\\ z(0)=0 \qquad z(1)=0\end{cases} \tag{11}$$

其中 $h(t,z)=g(t,z+7+2t)$. 详细地表示 h 为

$$h(t,z)=4[(z+7+2t)^2+3-(3+2t)^2+(z+7+2t)(3+2t)]$$

我们可以解 z 的边值问题(11)，从方程

$$y(t)=z(t)+7+2t$$

得到边值问题(10)的解. 然后从

$$x(t)=y\left(\frac{t-3}{2}\right)$$

得到边值问题(9)的解. ■

例 5 假如对边值问题(11)计算解 $z(0.5)=8.2$. 试问边值问题(9)对应的解是多少?

解 显然，$y(0.5)=z(0.5)+7+2(0.5)=16.2$ 并且 $0.5=(t-3)/2$ 推出 $t=4$. 答案是

$$x(4)=y(0.5)=z(0.5)+8=16.2$$ ■

定理 4(边值问题唯一解定理) 设 f 为 (t,s) 的连续函数，其中 $0\leqslant t\leqslant 1$，$-\infty<s<+\infty$. 假如在这个区域上

$$|f(t,s_1)-f(t,s_2)|\leqslant k|s_1-s_1| \qquad (k<8)$$

577

则两点边值问题

$$\begin{cases}x''=f(t,x)\\ x(0)=x(1)=0\end{cases} \tag{12}$$

在 $C[0,1]$ 中有唯一解.

证明 (概略地) 借助于习题 8.7.10，我们知道边值问题(12)等价于积分方程

$$x(t)=\int_0^1 G(t,s)f(s,x(s))\mathrm{d}s \tag{13}$$

其中 G 是习题 8.7.6 的格林函数. 积分方程(13)具有形式 $x=F(x)$，其中 F 是由积分定义的运算. 利用空间 $C[0,1]$中的 Banach 压缩映射定理，我们得到 F 有唯一的不动点，并且方程(12)和(13)有唯一解. ■

例 6 说明下列问题有唯一解:

$$\begin{cases}x''=2\exp(t\cos x)\\ x(0)=x(1)=0\end{cases} \tag{14}$$

解 这里 $f(t,s)=2\exp(t\cos s)$，由中值定理，得

$$|f(t,s_1)-f(t,s_2)|=\left|\frac{\partial f}{\partial s}(t,s_3)\right||s_1-s_2|$$

这里所需的导数满足关系

$$\left|\frac{\partial f}{\partial s}\right|=|2\exp(t\cos s)(-t\sin s)|\leqslant 2\mathrm{e}<5.437<8$$

由定理 4，两点边值问题(14)有唯一解. ■

习题 8.7

1. 求解两点边值问题 $x''=x$，$x(0)=1$，$x(1)=1$.
2. 确定使问题 $x''=x$，$x(0)=\alpha$，$x(1)=\beta$ 有解的所有数对(α,β).
3. (续)对问题 $x''=-x$，$x(0)=\alpha$，$x(\pi)=\beta$ 重复上题.
4. (续)给出类型(3)的两点边值问题存在多于一个解的例子. 提示：考虑上题.

578

5. 求解问题 $x''-2x'+x=0$，$x(0)=\alpha$，$x(1)=\beta$. 是否存在问题无解的数对(α,β)?
6. 证明两点边值问题 $x''=f(t)$，$x(0)=x(1)=0$ 是用公式

$$x(t)=\int_0^1 G(t,s)f(s)\mathrm{d}s$$

求解的，其中

$$G(t,s)=\begin{cases}s(t-1) & 0\leqslant s\leqslant t\leqslant 1\\ t(s-1) & 0\leqslant t\leqslant s\leqslant 1\end{cases}$$

函数 G 称为这个问题的格林函数.

7. 考虑下列两点边值问题.

i. $\begin{cases}x''=f(t,x,x')\\ x(a)=\alpha \qquad x(b)=\beta\end{cases}$

ii. $\begin{cases}x''=h^2 f(a+th,x,h^{-1}x')\\ x(0)=\alpha \qquad x(1)=\beta\end{cases}$

证明：若 x 是边值问题 ii 的一个解，则函数 $y(t)=x((t-a)/h)$是边值问题 i 的解，其中 $h=b-a$.

8. 用边值问题

$$\begin{cases}x''=t+x^2-3x'\\ x(3)=\alpha \qquad x(7)=\beta\end{cases}$$

说明定理 2. 这个问题等价于

$$\begin{cases}x''=48+64t+16x^2-12x'\\ x(0)=\alpha \qquad x(1)=\beta\end{cases}$$

9. 利用定理 4 证明两点边值问题

$$\begin{cases}x''=\cos(tx)\\ x(0)=1 \qquad x(1)=4\end{cases}$$

有唯一解.

10. (续)证明边值问题

$$\begin{cases}x''=f(t,x)\\ x(0)=0 \qquad x(1)=0\end{cases}$$

等价于积分方程

$$x(t)=\int_0^1 G(t,s)f(s,x(s))\mathrm{d}s$$

其中 G 是上面习题 8.7.6 的格林函数.

11. 证明下列两点边值问题有唯一解：

$$\begin{cases}x''=(t^3+5)x+\sin t\\ x(0)=x(1)=0\end{cases}$$

12. 证明下列问题有唯一解：

$$\begin{cases} x'' = \tan^{-1}x + 2x + \cos t \\ x(0) = x(1) = 0 \end{cases}$$

579

13. 求解边值问题

$$\begin{cases} x'' = x^2 \\ x(0) = 2/3 \qquad x(1) = 3/8 \end{cases}$$

14. (续)当边界条件简化为 $x(0)=0$，$x(1)=1$ 时求解上题. 参见 Davis[1962].

15. 求解边值问题

$$\begin{cases} x'' = x^3 \\ x(0) = 1 \qquad x(1) = 2-\sqrt{2} \end{cases}$$

16. 证明：若 x 是微分方程 $x''=f(x)$的解，则 $y(t)=x(t+c)$也是解.

17. 证明下列问题有唯一解：

$$\begin{cases} x'' = \dfrac{1}{2}\exp\{\dfrac{1}{2}(t+1)\cos(x+7-3t)\} \qquad (-1 \leqslant t \leqslant 1) \\ x(-1) = -10 \qquad x(1) = -4 \end{cases}$$

18. 判定下列形式的两点边值问题

$$\begin{cases} z'' = h(s,z) \\ z(0) = z(1) = 0 \end{cases}$$

等价于问题

$$\begin{cases} x'' = \cos(x^2t^2) \\ x(3) = 5 \qquad x(7) = 12 \end{cases}$$

19. 假定 v 和 w 是具有下列性质的 t 的已知函数：

$$\begin{cases} v'' = \cos t + ve^t + v't^2 \qquad v(1) = 5 \quad v(3) = 42 \\ w'' = \cos t + we^t + w't^2 \quad w(1) = 8 \quad w(3) = 9 \end{cases}$$

试问什么函数是下列两点边值问题的解？

$$\begin{cases} x'' = \cos t + xe^t + x't^2 \\ x(1) = 7 \quad x(3) = 20 \end{cases}$$

20. (多重选择题)

a. 两点边值问题

$$\begin{cases} x'' = -4x \\ x(0) = 1 \qquad x(\pi/2) = -1 \end{cases}$$

是具有下列性质的例子：

i. 无解

ii. 恰好两个解

iii. 恰好一个解

iv. 无初等函数解

v. 多于一个解

580

b. 对两点边值问题

$$\begin{cases} x'' = -4x \\ x(0) = 1 \qquad x(\pi/2) = 2 \end{cases}$$

重复 a.

21. 考虑两点边值问题

$$\begin{cases} x'' = f(t,x) \\ x(0) = x(1) = 0 \end{cases}$$

对下列什么函数 f，我们能断言存在唯一解？

ⅰ. $f(t,\ x)=t^2(1+x^2)^{-1}$

ⅱ. $f(t,\ x)=t\sin x$

ⅲ. $f(t,\ x)=(\tan t)(\tan x)$

ⅳ. $f(t,\ x)=t/x$

ⅴ. $f(t,\ x)=x^{1/3}$

22. 考察两点边值问题

$$\begin{cases} x'' = f(t,x) \\ x(0) = x(1) = 0 \end{cases}$$

对下列什么函数 f，我们能断言存在唯一解？

ⅰ. $f(t,\ x)=t^2 x^3$

ⅱ. $f(t,\ x)=t(\tan^{-1} x)$

ⅲ. $f(t,\ x)=t\, x^{4/3}$

ⅳ. $f(t,\ x)=\log(1+x^2)$

ⅴ. $f(t,\ x)=|tx|$

23. 证明：若 $x(t)$ 是

$$\begin{cases} x''(t) = f(t,x(t)) \\ x(a) = \alpha \qquad x(b) = \beta \end{cases}$$

的一个解，则 $y(s)=x(a+(b-a)s)$ 是

$$\begin{cases} y''(s) = (b-a)^2 f(a+(b-a)s, y(s)) \\ y(0) = \alpha \qquad y(1) = \beta \end{cases}$$

的解，因此 $z(s)=y(s)-[\alpha+(\beta-\alpha)\ s]$ 是

$$\begin{cases} z''(s) = (b-a)^2 f(a+(b-a)s, z(s)+\alpha+(\beta-\alpha)s) \\ z(0) = 0 \qquad z(1) = 0 \end{cases}$$

的解.

8.8 边值问题：打靶法

考虑两点边值问题

581

$$\begin{cases} x'' = f(t,x,x') \\ x(a) = \alpha \qquad x(b) = \beta \end{cases} \tag{1}$$

处理这个问题的一个自然的方法是用一个猜测值作为近似的初值 $x'(a)$，求解有关的初值问题. 然后可以积分该方程得到一个近似解，希望 $x(b)=\beta$. 若 $x(b)\neq\beta$，则可以改变 $x'(a)$的猜测值，接着再尝试. 这个过程称为**打靶**，有一些系统地做此项工作的方法.

用 z 表示 $x'(a)$的猜测值，使得相应的初值问题是

$$\begin{cases} x'' = f(t,x,x') \\ x(a) = \alpha \qquad x'(a) = z \end{cases} \tag{2}$$

用 x_z 表示这个初值问题的解. 目标是选择 z 使得 $x_z(b)=\beta$. 取

$$\phi(z) \equiv x_z(b) - \beta$$

使得我们的目标简化为对 z 求解方程 $\phi(z)=0$. 在此可应用第 3 章中考虑的求解单个非线性方程的方法. 例如，除了泛函迭代之外，两分法、割线法和牛顿法都可以利用. 因为 $\phi(z)$ 的每个值是通过数值求解一个初值问题来得到的，所以函数 ϕ 的计算量是非常大的.

8.8.1 割线法

回顾割线法是怎样应用于求解方程 $\phi(z)=0$ 的. 用 $\phi(z)$ 的两个值，譬如 $\phi(z_1)$ 和 $\phi(z_2)$，假设 $\phi(z)$ 是线性的. 利用通过点 $(z_2, \phi(z_2))$ 和 $(z_1, \phi(z_1))$ 的直线方程，我们有

$$\phi(z) - \phi(z_2) = \left(\frac{\phi(z_1) - \phi(z_2)}{z_1 - z_2}\right)(z - z_2)$$

若选择 z_3 使得 $\phi(z_3)=0$，则得到公式

$$z_3 = z_2 - \left(\frac{z_2 - z_1}{\phi(z_2) - \phi(z_1)}\right)\phi(z_2) \tag{3}$$

通过

$$z_n = z_{n-1} - \left(\frac{z_{n-1} - z_{n-2}}{\phi(z_{n-1}) - \phi(z_{n-2})}\right)\phi(z_{n-1})$$

可重复这个过程得到一连串的值 z_1，z_2，…，z_n. 这个式子定义了割线法.（见 3.3 节.）当已经得到 z 的若干个值使得 $\phi(z)$ 几乎为 0 时，我们停止这个过程并利用多项式插值去估计较好的值. 这里给出了怎样做的方法：假如 $\phi(z_1)$，$\phi(z_2)$，…，$\phi(z_n)$ 很小. 求一个插值表

$\phi(z_1)$	$\phi(z_2)$	…	$\phi(z_n)$
z_1	z_2	…	z_n

的多项式 p. 因此，多项式 p 有性质 $p(\phi(z_i))=z_i$，$1\leqslant i\leqslant n$. 下一个估计 z_{n+1} 是由等式 $p(0)=z_{n+1}$ 确定的. 这个过程相当于用多项式 p 逼近 ϕ 的**反函数**. 它的成功依赖于 ϕ 在根的一个邻域内有一个可微的反函数. 反过来这需要假定方程的根是单的. 582

8.8.2 线性函数

上面概述的打靶法在计算上的代价可能十分昂贵，所以考虑经济的方法是很重要的. 显然应该利用关于 $x'(a)$ 的校正值的任何部分信息. 因为高精度在打靶法的第一步基本上是被浪费的，所以用大步长求解初值问题也是可能的. 仅当 $\phi(z)$ 几乎是 0 时才应该使用小步长.

有一类问题，对它使用割线法一步就得到精确解. 事实上，当 ϕ 是线性函数时，以及当微分方程是线性的时候出现这种情况. 在线性情况下，两点边值问题具有形式

$$\begin{cases} x'' = u(t) + v(t)x + w(t)x' \\ x(a) = \alpha \qquad x(b) = \beta \end{cases} \tag{4}$$

下面将假定函数 u，v 和 w 在区间 $[a, b]$ 上是连续的. 假设我们已经用两个不同的初始条件两次求解(4)中的微分方程，得到解 x_1 和 x_2，

$$\begin{cases} x_1(a) = \alpha & x'_1(a) = z_1 \\ x_2(a) = \alpha & x'_2(a) = z_2 \end{cases} \tag{5}$$

现在形成 x_1 和 x_2 的一个线性组合：

$$y(t) = \lambda x_1(t) + (1-\lambda)x_2(t) \tag{6}$$

其中 λ 是一个参数. 容易验证 y 是微分方程的解并满足两个边界条件中的第 1 个，即 $y(a)=\alpha$. 我们选择 λ 使得 $y(b)=\beta$. 于是，

$$\beta = y(b) = \lambda x_1(b) + (1-\lambda)x_2(b)$$

并且

$$\lambda = \frac{\beta - x_2(b)}{x_1(b) - x_2(b)} \tag{7}$$

对线性问题(4)在计算机中实现这个想法时，我们可同时得到 x_1 和 x_2. 因此，可指定求解的两个初值问题为

[583]

$$\begin{cases} x'' = f(t,x,x') \\ x(a) = \alpha \quad x'(a) = 0 \end{cases} \qquad \begin{cases} x'' = f(t,x,x') \\ x(a) = \alpha \quad x'(a) = 1 \end{cases}$$

其中 $f(t, x, x') = u(t) + v(t)x + w(t)x'$. 第 1 个解是 x_1，第 2 个解是 x_2. 为产生一个 t 不显式出现的一阶方程组，我们再定义 $x_0 = t$，$x_3 = x'_1$，$x_4 = x'_2$. 因而带有初值的微分方程组是

$$\begin{cases} x_0 = 1 & x_0(a) = a \\ x'_1 = x_3 & x_1(a) = \alpha \\ x'_2 = x_4 & x_2(a) = \alpha \\ x'_3 = f(x_0, x_1, x_3) & x_4(a) = 0 \\ x'_4 = f(x_0, x_2, x_4) & x_5(a) = 1 \end{cases} \tag{8}$$

为了求解初值问题，应该把这个方程组输入到计算机程序中. 对 $a = t_0 \leqslant t_i \leqslant t_m = b$，离散函数的近似值 $x_1(t_i)$ 和 $x_2(t_i)$ 应该作为一维数组存放在计算机内存中. 其次，λ 的值应该用(7)式计算. 最后，根据(6)式在每个要求的 t 值上算出解 y.

定理 1(线性两点边值问题第一定理) *若线性两点边值问题(4)有解，则不是 x_1 本身是一个解就是 $x_1(b) - x_2(b) \neq 0$(且 y 是一个解).*

证明 设 y_0，y_1 和 y_2 是下列这些初值问题

$$\begin{array}{lll} y''_0 = u + vy_0 + wy'_0 & y_0(a) = \alpha & y'_0(a) = 0 \\ y''_1 = vy_1 + wy'_1 & y_1(a) = 1 & y'_1(a) = 0 \\ y''_2 = vy_2 + wy'_2 & y_2(a) = 0 & y'_2(a) = 1 \end{array}$$

的解. 由二阶线性微分方程的理论(特别地，由下面的定理 3)，(4)中的微分方程的通解是

$$y_0 + c_1 y_1 + c_2 y_2$$

其中 c_1 和 c_2 是任意的常数. 我们直接地看到(5)中的函数 x_1 和 x_2 是通解的特殊情况. 它们由下式给出

$$x_1 = y_0 + z_1 y_2 \qquad x_2 = y_0 + z_2 y_2 \tag{9}$$

因为已经假定(4)中的问题有解，存在 c_1 和 c_2 使得

$$\alpha = y_0(a) + c_1 y_1(a) + c_2 y_2(a)$$
$$\beta = y_0(b) + c_1 y_1(b) + c_2 y_2(b)$$

这些式子的第一个化简为 $c_1=0$，于是我们推断 c_2 的存在性，使得

$$\beta = y_0(b) + c_2 y_2(b) \tag{10}$$

若 $x_1(b)-x_2(b)\neq 0$，则由(6)式和(7)式定义的函数 y 是(4)的解. 若 $x_1(b)-x_2(b)=0$，则由(9)式有 $y_2(b)=0$. (10)式告诉我们 $y_0(b)=\beta$，而(9)式告诉我们 x_1 是一个解. ■ [584]

8.8.3 牛顿方法

我们转到方程(1)的更一般的(非线性)两点边值问题并考虑如何应用牛顿方法. 记得 x_z 被定义为问题

$$\begin{cases} x''_z = f(t, x_z, x'_z) \\ x_z(a) = \alpha \qquad x'_z(a) = z \end{cases} \tag{11}$$

的解. 我们要选择 z 使得

$$\phi(z) \equiv x_z(b) - \beta = 0$$

关于函数 ϕ 的牛顿公式是

$$z_{n+1} = z_n - \frac{\phi(z_n)}{\phi'(z_n)} \tag{12}$$

为确定 ϕ'，我们对(11)中的所有式子关于 z 求偏导数

$$\begin{cases} \dfrac{\partial x''_z}{\partial z} = \dfrac{\partial f}{\partial t}\dfrac{\partial t}{\partial z} + \dfrac{\partial f}{\partial x_z}\dfrac{\partial x_z}{\partial z} + \dfrac{\partial f}{\partial x'_z}\dfrac{\partial x'_z}{\partial z} \\ \dfrac{\partial}{\partial z}x_z(a) = 0 \qquad \dfrac{\partial}{\partial z}x'_z(a) = 1 \end{cases} \tag{13}$$

通过化简和引入新变量 $v=\partial x_z/\partial z$，上式变成

$$\begin{cases} v'' = f_{x_z}(t, x_z, x'_z)v + f_{x'_z}(t, x_z, x'_z)v' \\ v(a) = 0 \qquad v'(a) = 1 \end{cases} \tag{14}$$

我们认可这组式子为一个初值问题. (14)式中的微分方程称为**第一变分方程**. 它可以连同(11)一起步进求解. 最终可利用 $v(b)$，并且有

$$v(b) = \frac{\partial x_z(b)}{\partial z} = \phi'(z)$$

这使我们能够利用(12)式的牛顿方法求 ϕ 的根.

8.8.4 多重打靶

打靶法的一个发展称为多重打靶. 这里基本的策略是把给定的区间[a, b]分成子区间并试图在小段中求解整体问题. 当区间刚好被分成两部分[a, c]和[c, b]时，我们来描述将要做的工作.

如前所述，原问题为

[585]

$$\begin{cases} x'' = f(t, x, x') \\ x(a) = \alpha \qquad x(b) = \beta \end{cases}$$

在两个子区间上，我们求解初值问题得到两个函数 x_1 和 x_2：

$$\begin{cases} x''_1 = f(t,x_1,x'_1) & x_1(a) = \alpha \quad x'_1(a) = z_1 \quad (a \leqslant t \leqslant c) \\ x''_2 = f(t,x_2,x'_2) & x_2(b) = \beta \quad x'_2(b) = z_2 \quad (c \leqslant t \leqslant b) \end{cases}$$

注意 z_1 和 z_2 是我们配置的参数．函数 x_1 仅在区间$[a, c]$上需要，而 x_2 仅仅在区间$[c, b]$上需要．x_2 的数值解将按递减 t 的方向进行．

现在调整参数 z_1 和 z_2 直到分段函数

$$x(t) = \begin{cases} x_1(t) & a \leqslant t \leqslant c \\ x_2(t) & c \leqslant t \leqslant b \end{cases}$$

变成问题的解．因此，需要 x 和 x' 在点 c 上的连续性：$x_1(c) - x_2(c) = 0$ 和 $x'_1(c) - x'_2(c) = 0$．通过灵活地选择 z_1 和 z_2，这两个条件是满足的．代表性地这可用 3.2 节中的 2 维牛顿方法来完成．

对 k 个子区间的多重打靶将涉及 k 个**子函数**，每个子函数通过数值求解一个初值问题得到．这 k 个子函数的初值构成一个有 $2k$ 个参数的集合．在区间的 $k-1$ 个内分点中的每一个点上，必须利用整体函数及其导数的连续性．这就提供 $2k-2$ 个条件．两个端点条件存在，所以条件个数匹配参数个数．最后得到的非线性方程组被迭代地求解，例如用高维牛顿方法．

可得到的两点边值问题的大多数软件是针对一阶常微分方程组

$$X' = F(t,X)$$

编写的，其中 $X = (x_1, x_2, \cdots, x_n)^T$，$F = (f_1, f_2, \cdots, f_n)^T$．边界条件经常允许是相当一般的条件．例如，某些代码允许边界条件为

$$G(X(a), X(b)) = 0$$

其中 $G = (g_1, g_2, \cdots, g_n)^T$．某些软件需要用户完成 F 的雅可比阵，它是元素为 $J_{ij} = \partial f_i / \partial x_j$ 的 $n \times n$ 矩阵．

例 1 假设对问题

$$\begin{cases} x'' = tx + \cos x' \\ x(3) + x'(5) = 7 \qquad x'(3)^2 x(5) = 10 \end{cases}$$

[586] 应用刚才所述类型的软件，试问 F，G 和 J 是什么？

解 取 $x_1 = x$，$x_2 = x'$．则问题可表述成

$$\begin{cases} x'_1 = x_2 & x_1(3) + x_2(5) - 7 = 0 \\ x'_2 = tx_1 + \cos x_2 & x_2(3)^2 x_1(5) - 10 = 0 \end{cases}$$

函数 F 和 G 可从这些式子中读出．雅可比函数由下式给出

$$J(t,X) = \begin{bmatrix} 0 & 1 \\ t & -\sin x_2 \end{bmatrix}$$

■

8.8.5 二阶线性方程

这里引用二阶线性微分方程一般理论的两个重要定理．这些内容的标准参考文献是 Coddington and Levinson[1955]．

定理 2(二阶线性微分方程第二定理) 若 u，v 和 w 是闭区间$[a, b]$上的连续函数，则对

任何实数对 α 和 α'，初值问题

$$\begin{cases} x'' = u + vx + wx' \\ x(a) = \alpha \qquad x'(a) = \alpha' \end{cases}$$

在$[a, b]$上有唯一解.

定理 3(二阶线性微分方程第三定理) 非齐次方程

$$x'' - vx - wx' = u \tag{15}$$

的每个解可表示成 $x_0 + c_1x_1 + c_2x_2$，其中 x_0 是(15)式的特解，而 x_1 和 x_2 构成齐次方程

$$x'' - vx - wx' = 0$$

的线性无关的解集.

习题 8.8

1. 利用真解，明确地求出下列情况中的函数 ϕ.

$$\begin{cases} x'' = -x \\ x(0) = 1 \qquad x\left(\frac{\pi}{2}\right) = 3 \end{cases}$$

2. 确定真解并明确地求出下列情况中的函数 ϕ.

$$\begin{cases} x'' = -(x')^2x^{-1} \\ x(1) = 3 \qquad x(2) = 5 \end{cases}$$

利用 ϕ 求解边值问题.

587

3. 解析求解三点边值问题：

$$\begin{cases} x''' = -e^t + 4(t+1)^{-3} \\ x(0) = -1 \quad x(1) = 3 - e + 2\ln 2 \quad x(2) = 6 - e^2 + 2\ln 3 \end{cases}$$

4. 说明如何利用打靶法求解下面类型的两点边值问题，其中常数 α，β 和 c_{ij} 已知：

$$\begin{cases} x'' = u(t) + v(t)x + w(t)x' \\ c_{11}x(a) + c_{12}x'(a) = \alpha \\ c_{21}x(b) + c_{22}x'(b) = \beta \end{cases}$$

提示：设 x_1 是具有特定的初始条件 $c_{11}x_x(a) + c_{12}x_1(a) = \alpha$ 的微分方程的解. 设 x_2 是具有特定的初始条件 $x_2(a) = -c_{12}$ 和 $x'_2(a) = c_{11}$ 的微分方程的解. 考虑 $x_1 + \lambda x_2$.

5. 对应于线性微分方程

$$x'' = a(t) + b(t)x + c(t)x'$$

的第一变分方程是什么？

6. 微分方程

$$x'' = \cos(tx) + \sin(t^2x')$$

的第一变分方程是什么？第一变分方程能否用它本身数值求解，或者它必须用 x''的方程一起来求解？

7. 证明：若 x_1 和 x_2 是具有初始条件 $x_1(a) = x_2(a) = \alpha$，$x'_1(a) = 0$ 和 $x'_2(a) = 1$ 的微分方程 $x'' = u(t) + v(t)x + w(t)x'$的解，则 $x_2 - x_1$ 是(12)式中第一变分问题的解.

8. 证明：若对 ϕ 用牛顿方法求解线性两点边值问题，并且用第一变分方程计算 ϕ'，则所得结果与(6)式和(7)式给出的结果相同.

9. 证明：在线性微分方程的情况中函数 ϕ 是线性的.

10. 求问题

$$\begin{cases} x'' = -9x \\ x(0) = 1 \qquad x\left(\frac{\pi}{6}\right) = 5 \end{cases}$$

的解．首先求问题

$$\begin{cases} x'' = -9x \\ x(0) = 1 \qquad x'(0) = z \end{cases}$$

的解 x_z，然后调整 z 使得 $x_z(\frac{\pi}{6})=5$．如果 $x(\frac{\pi}{3})=5$，请描述结果将如何改变？

11. 明确地求出下列情况中的函数 ϕ.

$$\begin{cases} x'' = -2t(x')^2 \\ x(0) = 1 \qquad x(1) = 1+\frac{\pi}{4} \end{cases}$$

利用 ϕ 求解边值问题.

588

12. 若 x_z 是初值问题 $x''=x$，$x(0)=0$，$x'(0)=z$ 的解，试问 $x_z(1)$是什么？提示：用 x 乘以微分方程并积分.
13. 对两点边值问题 $x''=x$，$x(0)=0$，$x(1)=17$，试问函数 $\phi(z)$是什么？
14. 若 x_z 是初值问题 $x''=-x$，$x(0)=5$，$x'(0)=z$ 的解，并且 $\phi(z)=x_z(\pi/2)-3$，则 $\phi'(z)$是什么？
15. 利用基于割线法的打靶法求解两点边值问题

$$\begin{cases} x'' - 37t^2 x' = 95 \\ x(6) = 1 \qquad x(12) = 2 \end{cases}$$

我们得到两个数对(z_i，$x_{z_i}(b)$)，$i=1$，2，比方说(4，5)和(2，9)．试问在这个过程中下一步迭代求解的是什么初值问题？

16. 查阅线性两点边值问题的讨论并证明 $c_1=0$ 和 $c_2=[\beta-y_0(b)]/y_2(b)$.
17. (续)证明：若 $y_2(b)=0$，则边值问题对一切(α，β)将不可解.
18. 证明：倘若解存在时，下列过程将可解两点边值问题(4)．解两个初值问题

$$\begin{cases} x_1'' = u + vx_1 + wx_1' \quad x_1(a) = \alpha \quad x_1'(a) = 0 \\ x_2'' = vx_2 + wx_2' \qquad x_2(a) = 0 \quad x_2' = 1 \end{cases}$$

则不是 x_1 本身是解就是 x_1+cx_2 是解，其中 c 是常数$[\beta-x_1(b)]/x_2(b)$.

19. 证明：问题 $x''=x$，$x(a)=\alpha$，$x(b)=\beta$ 总有唯一解．(假设 $a\neq b$).

计算机习题 8.8

1. 编写出打靶法的算法．用打靶法求解两点边值问题

$$\begin{cases} x'' = e^t + x\cos t - (t+1)x' \\ x(0) = 1 \qquad x(1) = 3 \end{cases}$$

注意问题是线性的．利用步长 $h=0.01$ 的四阶龙格-库塔方法.

2. 编写课本中列举的求解线性两点边值问题的算法．查阅(4)～(8)式．编写基于这个算法的一个通用代码．函数 u，v 和 w 应该由用户提供子程序形式的代码.
3. (续)对下面例子

$$\begin{cases} u(t) = e^{t-3}, v(t) = t^2+2, w(t) = \sin t \\ a = 2.6, b = 5.1, \alpha = 7.0, \beta = -3 \end{cases}$$

测试上题中的代码.

8.9 边值问题：有限差分法

两点边值问题的另一种方法由 t 区间的初始离散化和导数的近似公式所构成. 下面的两个公式是特别有用的： 589

$$x'(t) = (2h)^{-1}[x(t+h) - x(t-h)] - \frac{1}{6}h^2 x'''(\xi) \tag{1}$$

$$x''(t) = h^{-2}[x(t+h) - 2x(t) + x(t-h)] - \frac{1}{12}h^2 x^{(4)}(\tau) \tag{2}$$

8.9.1 二阶微分方程

假定求解的问题是

$$\begin{cases} x'' = f(t,x,x') \\ x(a) = \alpha \qquad x(b) = \beta \end{cases} \tag{3}$$

设区间$[a, b]$用点 $a=t_0$，t_1，t_2，…，$t_{n+1}=b$ 分割. 这些点不一定是等距的，但是实际上它们往往是等距的. 当然，若这些点不是均匀分布的话，则必定引入(1)和(2)的更复杂的形式. 故为简单起见，我们假设

$$t_i = a + ih \qquad 0 \leqslant i \leqslant n+1 \qquad h = (b-a)/(n+1) \tag{4}$$

用 y_i 表示 $x(t_i)$的近似值. 于是，(3)的离散形式是

$$\begin{cases} y_0 = \alpha \\ h^{-2}(y_{i-1} - 2y_i + y_{i+1}) = f(t_i, y_i, (2h)^{-1}(y_{i+1} - y_{i-1})) \qquad (1 \leqslant i \leqslant n) \\ y_{n+1} = \beta \end{cases} \tag{5}$$

8.9.2 线性情况

在(5)式中，未知量是 y_1，y_2，…，y_n，而求解的是 n 个方程. 若 f 以非线性方式包含 y_i，则这些方程是非线性的，而且一般说来求解很困难. 然而，我们假定 f 关于 x 和 x'是线性的. 则它具有下列形式

$$f(t,x,x') = u(t) + v(t)x + w(t)x' \tag{6}$$

现在方程组(5)是一个线性方程组，它可写成下列形式

$$\begin{cases} y_0 = \alpha \\ \left(-1-\frac{1}{2}hw_i\right)y_{i-1} + (2+h^2v_i)y_i + \left(-1+\frac{1}{2}hw_i\right)y_{i+1} = -h^2u_i \quad (1 \leqslant i \leqslant n) \\ y_{n+1} = \beta \end{cases} \tag{7}$$

我们已经记 $u_i=u(t_i)$，$v_i=v(t_i)$等等.

下面，我们引进缩写

590

$$a_i = -1 - \frac{1}{2}hw_{i+1}$$

$$d_i = 2 + h^2 v_i$$

$$c_i = -1 + \frac{1}{2}hw_i$$

$$b_i = -h^2 u_i$$

使得方程组看上去像这样：

$$\begin{bmatrix} d_1 & c_1 & & & & \\ a_1 & d_2 & c_2 & & & \\ & a_2 & d_3 & c_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & a_{n-2} & d_{n-1} & c_{n-1} \\ & & & & a_{n-1} & d_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix} = \begin{bmatrix} b_1 - a_0\alpha \\ b_2 \\ b_3 \\ \vdots \\ b_{n-1} \\ b_n - c_n\beta \end{bmatrix} \tag{8}$$

因为没有标明的元素为 0，所以这个方程组是三对角的，可用特殊的高斯算法求解．注意，如果 h 很小且 $v_i>0$，则方程组的矩阵是对角占优的，因为

$$|2+h^2v_i| > \left|1+\frac{1}{2}hw_i\right| + \left|1-\frac{1}{2}hw_i\right| = 2$$

这里必须假定 $|\frac{1}{2}hw_i|\leqslant 1$，因为这样两项 $1\pm\frac{1}{2}hw_i$ 都是非负的．因此，我们假定 $v_i>0$ 以及 h 足够小使不等式 $|\frac{1}{2}hw_i|<1$ 成立．

后面将需要下列等式

$$\begin{aligned} |d_i|-|c_i|-|a_{i-1}| &= 2+h^2v_i-(1-\frac{1}{2}hw_i)-(1+\frac{1}{2}hw_i) \\ &= h^2v_i \end{aligned} \tag{9}$$

8.9.3 收敛性

我们将着手证明当 h 收敛于 0 时，离散解收敛于边值问题的解．为知道边值问题

$$\begin{cases} x'' = u + vx + wx' \\ x(a) = \alpha \qquad x(b) = \beta \end{cases} \tag{10}$$

有唯一解，我们引用 Keller[1968，第 9 页]的一个定理，在 u，v 和 w 属于 $C[a, b]$ 且 $v>0$ 的假设之下得出这个结论．所以采用这些假设．Keller 的定理如下．

定理 1（边值问题解的存在性和唯一性定理） 边值问题

$$\begin{cases} x'' = f(t,x,x') \\ c_{11}x(a) + c_{12}x'(a) = c_{13} \\ c_{21}x(b) + c_{22}x'(b) = c_{23} \end{cases}$$

591 *在区间 $[a, b]$ 上有唯一解，倘若：*

1. f 及其一阶偏导数 f_t，f_x 和 $f_{x'}$ 在域 $D=[a, b]\times\mathbb{R}\times\mathbb{R}$ 上连续．

2. $f_x>0$，$|f_x|\leqslant M$ 且在 D 上 $|f_{x'}|\leqslant M$．

3. $|c_{11}|+|c_{12}|>0$，$|c_{21}|+|c_{22}|>0$，$|c_{11}|+|c_{21}|>0$，且 $c_{11}c_{12}\leqslant 0\leqslant c_{21}c_{22}$．

用 $x(t)$ 表示问题的真解，且用 y_i 表示离散问题的解．注意 y_i 依赖于 h．我们将估计 $|x_i-y_i|$ 并指出当 $h\to 0$ 时它收敛于 0．当然，这里 x_i 表示 $x(t_i)$．

借助于公式(1)和(2)，我们看到对于 $1\leqslant i\leqslant n$，$x(t)$ 满足下列方程组：

$$h^{-2}(x_{i-1}-2x_i+x_{i+1}) - \frac{1}{12}h^2x^{(4)}(\tau_i) \tag{11}$$

$$= u_i + v_i x_i + w_i\left[(2h)^{-1}(x_{i+1} - x_{i-1}) - \frac{1}{6}h^2 x'''(\xi_i)\right] \tag{12}$$

另一方面，离散解满足等式

$$h^{-2}(y_{i-1} - 2y_i + y_{i+1}) = u_i + v_i y_i + w_i(2h)^{-1}(y_{i+1} - y_{i-1})$$

如果从(11)式中减去(12)式，并记 $e_i \equiv x_i - y_i$，则结果是

$$h^{-2}(e_{i-1} - 2e_i + e_{i+1}) = v_i e_i + w_i(2h)^{-1}(e_{i+1} - e_{i-1}) + h^2 g_i \tag{13}$$

其中

$$g_i = \frac{1}{12}x^{(4)}(\tau_i) - \frac{1}{6}x'''(\xi_i)$$

合并同类项并用 $-h^2$ 相乘以后，得到一个类似于(7)式的等式：

$$\left(-1 - \frac{1}{2}hw_i\right)e_{i-1} + (2 + h^2 v_i)e_i + \left(-1 + \frac{1}{2}hw_i\right)e_{i+1} = -h^4 g_i \tag{14}$$

利用前面引进的系数，我们记上式为

$$a_{i-1}e_{i-1} + d_i e_i + c_i e_{i+1} = -h^4 g_i$$

设 $\lambda = \| e \|_\infty$ 并选择一个指标 i 使得

$$| e_i | = \| e \|_\infty = \lambda$$

这里 e 是向量 $e = (e_1, e_2, \cdots, e_n)$. 然后从(14)式，我们得到

$$| d_i || e_i | \leqslant h^4 | g_i | + | c_i || e_{i+1} | + | a_{i-1} || e_{i-1} |$$

利用(9)式，我们得到

$$| d_i | \lambda \leqslant h^4 \| g \|_\infty + | c_i | \lambda + | a_{i-1} | \lambda$$

$$\lambda(| d_i | - | c_i | - | a_{i-1} |) \leqslant h^4 \| g \|_\infty$$

$$h^2 v_i \lambda \leqslant h^4 \| g \|_\infty$$

$$\| e \|_\infty \leqslant h^2 [\| g \|_\infty / \inf v(t)]$$

由(13)式，$\| g \|_\infty \leqslant \| x^{(4)} \|_\infty / 12 + \| x''' \|_\infty / 6$. 表达式 $\| g \|_\infty / \inf v(t)$ 是一个与 h 无关的界. 因此，我们看到当 $h \to 0$ 时 $\| e \|_\infty$ 是 $\mathcal{O}(h^2)$. 592

习题 8.9

1. 解两点边值问题

$$\begin{cases} x'' + 2x' + 10t = 0 \\ x(0) = 1 \qquad x(1) = 2 \end{cases}$$

利用 $h = \frac{1}{2}$ 的有限差分法求 $x(\frac{1}{2})$.

2. 为了适应下列形式的边界条件

$$c_{11}x(a) + c_{12}x'(a) + c_{13} = c_{21}x(b) + c_{22}x'(b) + c_{23} = 0$$

请指出在线性方程组(7)中必须作怎样的修正.

3. 倘若系数函数是连续的且 $v > 0$，利用 Keller 定理证明边值问题(10)有唯一解.

计算机习题 8.9

1. 编写一个用课本上描述的有限差分法求解线性两点边值问题的通用的计算机程序. 允许用户提供 a，α，b，β，n 以及像(3)，(4)和(6)式那样的函数 u，v 和 w.

2.（续）对下面的例子测试上题中编写的程序：

a. $\begin{cases} x''=-x \\ x(0)=3 \qquad x\left(\frac{\pi}{2}\right)=7 \end{cases}$

b. $\begin{cases} x''=2e^t-x \\ x(0)=2 \qquad x(1)=e+\cos 1 \end{cases}$

此外，计算这两种测试情况中数值解的误差．解分别是：a. $x(t)=7\sin t+3\cos t$. b. $x(t)=e^t+\cos t$.

8.10 边值问题：配置法

配置法提供一个策略，利用它可以处理应用数学中的许多问题．首先，我们给出一个一般的描述．假定我们有一个线性算子 L（例如，积分算子或微分算子），并且希望求解方程

$$Lu = w \tag{1}$$

在此方程中，w 已知而 u 是要求的．若干个求解方程(1)的近似方法从选取某个基向量组 $\{v_1, v_2, \cdots, v_n\}$ 开始，然后用下列形式的向量

$$u = c_1v_1 + c_2v_2 + \cdots + c_nv_n \tag{2}$$

尝试去解方程(1)．因为 L 是线性算子，所以我们有

593

$$Lu = \sum_{j=1}^{n} c_jLv_j$$

并且由方程(1)得到

$$\sum_{j=1}^{n} c_jLv_j = w \tag{3}$$

一般说来，我们不能求解方程组(3)中的系数 $c_1, c_2, \cdots, c_n$. 但是或许可使方程(3)几乎成立．在配置法中，向量 u, w 和 v_j 都是公共定义域上的函数．然后，我们要求函数 w 和 $\sum_{j=1}^{n} c_jLv_j$ 在 n 个给定点上的值相同：

$$\sum_{j=1}^{n} c_j(Lv_j)(t_i) = w(t_i) \qquad (1 \leqslant i \leqslant n) \tag{4}$$

这是一个 n 个线性方程的方程组，由这个方程组，我们可以计算 n 个未知系数 c_j 的值．当然，应该选择函数 v_j 和点 t_i 使元素为 $(Lv_j)(t_i)$ 的矩阵非奇异．

8.10.1 施图姆-刘维尔边值问题

现在考虑这个方法如何用于施图姆-刘维尔两点边值问题：

$$\begin{cases} u'' + pu' + qu = w \\ u(0) = 0 \qquad u(1) = 0 \end{cases} \tag{5}$$

这里函数 p, q 和 w 都是已知的，并假定在区间 $[0, 1]$ 上连续．函数 u 是未知的，它也定义在 $[0, 1]$ 上，但我们期望它是二次连续可微的．如果取

$$Lu \equiv u'' + pu' + qu \tag{6}$$

来定义算子 L，则这个问题就变成前面列举的格式中的一个例子．我们在向量空间

$$V = \{u \in C^2[0,1] : u(0) = u(1) = 0\} \tag{7}$$

中寻找一个解.

因此，如果从 V 中选择一组基函数 $\{v_1, v_2, \cdots, v_n\}$，则齐次边界条件将自动满足. 提出它本身是(双指标)集的一组函数

$$v_{jk}(t) = t^j(1-t)^k \qquad (j \geqslant 1, k \geqslant 1) \tag{8}$$

容易验证

$$v'_{jk} = jv_{j-1,k} - kv_{j,k-1} \tag{9}$$

$$v''_{jk} = j(j-1)v_{j-2,k} - 2jkv_{j-1,k-1} + k(k-1)v_{j,k-2} \tag{10}$$

根据(9)式和(10)式，写出函数 Lv_{jk} 是一件简单的事情. 如果从(8)式中所述的集合中选择 n 个函数，并且在 $[0, 1]$ 中选择 n 个点 t_i，则可尝试求解配置方程(4)并得到问题(5)的一个近似解.

594

8.10.2 三次 B 样条

或许对这样的问题来说，基函数的一种较好的选择是一组 B 样条. 在描述如何利用 B 样条的优势中，我们取模型问题为稍微更一般的情况：

$$\begin{cases} u'' + pu' + qu = w \\ u(a) = \alpha \qquad u(b) = \beta \end{cases} \tag{11}$$

为了使基函数具有两次连续导数，我们只考虑 B 样条 B_i^k，$k \geqslant 3$. 为简单起见，取 $k=3$. 另外，取等距结点：$t_{i+1} - t_i = h$. 最后，我们将利用结点作为配置点. 设 n 是使用的基函数的个数(以及要确定的系数的个数). 为确定 n 个系数，应该存在总数为 n 的条件. 因为存在两个端点条件，即

$$\sum_{j=1}^{n} c_j v_j(a) = \alpha \quad 和 \quad \sum_{j=1}^{n} c_j v_j(b) = \beta \tag{12}$$

所以我们看到应该有 $n-2$ 个配置条件：

$$\sum_{j=1}^{n} c_j (Lv_j)(t_i) = w(t_i) \qquad (1 \leqslant i \leqslant n-2) \tag{13}$$

这些考虑导致我们定义

$$h = (b-a)/(n-3) \tag{14}$$

$$t_i = a + (i-1)h \qquad (i = 0, \pm 1, \pm 2, \cdots) \tag{15}$$

位于 $[a, b]$ 中的结点 t_i 是 $a = t_1, t_2, \cdots, t_{n-2} = b$（这是容易验证的）. 这些结点都是配置点. 为定义 B 样条 B_j^3，需要某些在区间 $[a, b]$ 之外的点. 结点的排列如图 8-4 所示.

图 8-4 结点的配置

三次 B 样条的重要性是它们在 $[a, b]$ 上不恒等于 0. 这些三次 B 样条是 B_{-2}^3，B_{-1}^3，B_0^3，B_1^3，$\cdots$，B_{n-3}^3. 因此，我们取 $v_j = B_{j-3}^3$，$1 \leqslant j \leqslant n$.

因为结点是等距的，所以函数 v_j 可从用 B 表示的单个 B 样条得到，B 的定义如下：

595

$$B(t)=\begin{cases}(t+2)^3/6 & \text{在}[-2,-1]\text{上}\\ [1+3(t+1)+3(t+1)^2-3(t+1)^3]/6 & \text{在}[-1,0]\text{上}\\ B(-t) & \text{在}[0,2]\text{上}\\ 0 & \text{其他}\end{cases} \tag{16}$$

因而，容易验证

$$v_j(t)=B\left(\frac{t-a}{h}-j+2\right) \tag{17}$$

函数 B 的图像显示在图 8-5 中．读者应该检验 B 确实是一个三次样条．

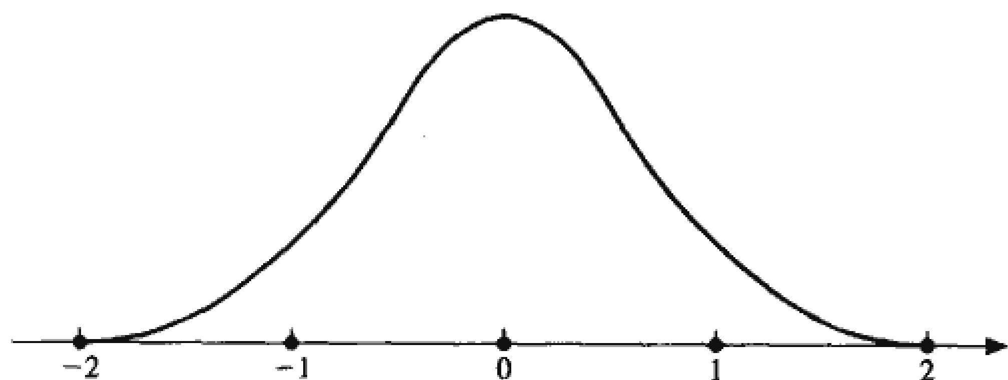


图 8-5 三次 B 样条

为计算 $(Lv_j)(t_i)$ 需要 v_j 的一阶和二阶导数．这些导数容易从(16)式和(17)式得到．在利用三次 B 样条执行配置法的计算机程序中，必须编写计算 $B(t)$，$B'(t)$ 和 $B''(t)$ 的子程序．因而，可有效地计算矩阵元素 $(Lv_j)(t_i)$．因为每个函数 v_j 有一个小的支撑，所以这是一个**带状矩阵**．打算作为产品使用的通用程序应该利用系数矩阵的稀疏性质．

计算机习题 8.10

a. 对用配置法求解下列形式的两点边值问题

$$\begin{cases}u''(t)+p(t)u'+q(t)u(t)=w(t)\\ u(a)=\alpha \qquad u(b)=\beta\end{cases}$$

编写一个通用的代码程序．用户应该详细说明：(1)函数 p，q 和 w；(2)实数 a，α，b 和 β，$a<b$；(3)所使用的基函数的个数 n；(4)基函数 v_i 和它们的导数 v_i' 和 v_i''．然后，计算机程序在区间 $[a,b]$ 中生成等距的 $n-2$ 个配置点 t_i，$1\leqslant i\leqslant n-2$，包括端点．接着，程序产生一个有 n 个未知数 c_1，c_2，…，c_n 的 n 个方程的线性方程组，如下所示：

$$\begin{cases}\sum_{j=1}^{n}c_j(Lv_j)(t_i)=w(t_i) \qquad (1\leqslant i\leqslant n-2)\\ \sum_{j=1}^{n}c_jv_j(a)=\alpha \qquad \sum_{j=1}^{n}c_jv_j(b)=\beta\end{cases}$$

这个程序称为计算系数 c_j 的线性方程解法器．最后，通过在 $[a,b]$ 中的 $2n-5$ 个等距点上求 u 的值和残差 $Lu-w$ 来检验近似解 $u=\sum_{j=1}^{n}c_jv_j$．这些点应该包括配置点(那里残差应该是 0)以及配置点的中点．

596

b. 通过求解

$$\begin{cases} u'' + (\sin t)u' + (t^2+2)u = e^{t-3} \\ u(2.6) = 7 \qquad u(5.1) = -3 \end{cases}$$

测试 a 中编写的程序．利用等距结点的三次 B 样条．配置点应该是$[a, b]$上的结点．

8.11 线性微分方程

这里我们考虑 n 个常系数的线性微分方程的方程组．假设方程组是**自控的**，这意味着独立变量 t 不显式地出现．这样的一个方程组可表示成

$$\begin{cases} x_1' = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ x_2' = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \quad\vdots \\ x_n' = a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n \end{cases} \tag{1}$$

使用向量矩阵记号，它简化为

$$X' = AX \tag{2}$$

其中 $X=(x_1, x_2, \cdots, x_n)^{\mathrm{T}}$，$X'=(x_1', x_2', \cdots, x_n')^{\mathrm{T}}$．

8.11.1 特征值和特征向量

我们尝试用 $X(t)=e^{\lambda t}V$ 这种形式的向量作为这个问题的解，其中 V 是一个常向量．把这个试验的解代入(2)式，我们得到

$$\lambda e^{\lambda t}V = e^{\lambda t}AV \tag{3}$$

因此，若

$$AV = \lambda V$$

则向量函数 $e^{\lambda t}V$ 确实是(2)式的一个解．我们已经证明了下列定理．

定理 1(线性微分方程的特征值和特征向量) *若 λ 是矩阵 A 的一个特征值，而 V 是相应的特征向量，则 $X(t)=e^{\lambda t}V$ 是方程 $X'=AX$ 的解．*

这个定理指出求解微分方程 $X'=AX$ 会涉及 A 的特征值和特征向量的某些知识．此外，相似矩阵的理论指出如何通过改变变量去简化线性微分方程组．这些不久将加以考虑． 597

在下面的定理中描述了方程 $X'=AX$ 的最令人高兴的情况．

定理 2(解空间的基定理) *若 $n\times n$ 矩阵 A 有一组线性无关的特征向量 $V_1, V_2, \cdots, V_n$，$AV_i=\lambda_i V_i$，则方程 $X'=AX$ 的解空间有一组基 $X_i=e^{\lambda_i t}V_i$，$1\leqslant i\leqslant n$.*

证明 因为 $\sum_{i=1}^{n} c_iX_i=0$ 导致 $\sum_{i=1}^{n} c_ie^{\lambda_i t}V_i=0$，$c_ie^{\lambda_i t}=0$ 和 $c_i=0$，$1\leqslant i\leqslant n$，所以向量组 $\{X_1, X_2, \cdots, X_n\}$线性无关．

为了证明问题中的一组向量张成 $X'=AX$ 的解空间，设 X 是任意解．初值向量 $X(0)$作为 $\mathbb{R}^n$或$\mathbb{C}^n$的一个元素，它是 $V_1, V_2, \cdots, V_n$ 的一个线性组合，例如

$$X(0) = \sum_{i=1}^{n} c_iV_i$$

定义 $Y=\sum_{i=1}^{n} c_iX_i$．则

$$Y' = \sum_{i=1}^{n} c_i X'_i = \sum_{i=1}^{n} c_i \lambda_i e^{\lambda_i t} V_i = \sum_{i=1}^{n} c_i e^{\lambda_i t} A V_i = A(\sum_{i=1}^{n} c_i X_i) = AY$$

于是，Y 和 X 都是微分方程的解，且它们有相同的初值：$Y(0)=X(0)$.（为什么?）由初值问题的唯一性定理，我们得到 $Y=X$，或换言之，$X=\sum_{i=1}^{n} c_i X_i$. ■

若 A 有定理 2 中提到的性质，则存在一个其列向量为 V_1，V_2，…，V_n 的非奇异矩阵 P，等式 $AV_i=\lambda_i V_i$，转换成矩阵记号为

$$AP = P\Lambda \tag{4}$$

其中 Λ 是在对角线上有 λ_1，λ_2，…，λ_n 的对角阵. 考虑用 $X=PY$ 表述的(相关)变量的改变. 因为 P 非奇异，所以我们可从 X 重新获得 Y. Y 有下列性质：

$$Y' = P^{-1}X' = P^{-1}AX = P^{-1}APY = \Lambda Y \tag{5}$$

因为 Λ 是对角阵，因此，关于 Y 的微分方程比关于 X 的微分方程简单得多. 在方程组 $Y'=\Lambda Y$ 中各自的方程是**非耦合的**，可以分别求解.

例 1 当

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} \qquad X(0) = \begin{bmatrix} 5 \\ 7 \\ 6 \end{bmatrix}$$

[598] 时求解初值问题 $X'=AX$.

解 矩阵 $A-\lambda I$ 是

$$\begin{bmatrix} 1-\lambda & 0 & 1 \\ 0 & -\lambda & 0 \\ 0 & 0 & -1-\lambda \end{bmatrix}$$

而它的行列式是 **A 的特征多项式**：

$$(1-\lambda)(-\lambda)(-1-\lambda)$$

这个多项式的根是 A 的特征值，它们是 $\lambda_1=1$，$\lambda_2=0$，$\lambda_3=-1$. 对每种情况，我们通过解 $AV_i=\lambda_i V_i$ 求特征向量. 把这些向量作为矩阵 P 的列向量，得到

$$P = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix}$$

其次，我们求

$$P^{-1} = \begin{bmatrix} 1 & 0 & \frac{1}{2} \\ 0 & 1 & 0 \\ 0 & 0 & -\frac{1}{2} \end{bmatrix}$$

若 $Y=(y_1, y_2, y_3)^T$，则关于 Y 的初值问题是 $Y'=\Lambda Y$，其中

$$\Lambda = P^{-1}AP = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

因此，我们有

$$\begin{cases} y_1' = y_1 \\ y_2' = 0 \\ y_3' = -y_3 \end{cases} \qquad Y(0) = P^{-1}X(0) = \begin{bmatrix} 8 \\ 7 \\ -3 \end{bmatrix}$$

而它的解是

$$y_1 = 8e^t \qquad y_2 = 7 \qquad y_3 = -3e^{-t}$$

因为 $X=PY$，所以 $X=(x_1, x_2, x_3)^T$ 对应的解是

$$x_1 = 8e^t - 3e^{-t} \qquad x_2 = 7 \qquad x_3 = 6e^{-t}$$

■

8.11.2 矩阵指数

有一个求解方程组 $X'=AX$ 的优美方法．这个方法在希望数值计算解之前不必涉及 A 的特征值．我们首先定义矩阵指数．

599

定义 1（矩阵指数的定义） 若 A 是一个方阵，我们取

$$e^A = I + A + \frac{1}{2!}A^2 + \frac{1}{3!}A^3 + \cdots \tag{6}$$

这个定义是从标准级数

$$e^z = 1 + z + \frac{1}{2!}z^2 + \frac{1}{3!}z^3 + \cdots \tag{7}$$

中用矩阵替代复变量 z 导出的．为了观察 e^A 的级数的收敛性，我们在$\mathbb{C}^n$上取任意范数并使用 $n\times n$ 矩阵相应的从属矩阵范数．级数的尾部可估计如下：

$$\left\| \sum_{k=m}^{\infty} \frac{1}{k!}A^k \right\| \leqslant \sum_{k=m}^{\infty} \frac{1}{k!} \| A^k \| \leqslant \sum_{k=m}^{\infty} \frac{1}{k!} \| A \|^k \tag{8}$$

当 $z=\|A\|$ 时，这个最后的表达式是通常的指数级数的尾部．因此，当 $m\to\infty$时，e^A 的级数的尾部收敛于 0．（众所周知，这个理由采用了给定范数的 $n\times n$ 矩阵空间的完备性.）

若 t 是一个实变量，则 $tA=At$，由我们的定义得出

$$e^{At} = \sum_{k=0}^{\infty} \frac{t^k}{k!}A^k \tag{9}$$

在级数中关于 t 的微分以及随后的简化给出

$$\frac{d}{dt}e^{At} = Ae^{At} \tag{10}$$

定理 3（初值问题的解的定理） 初值问题

$$\begin{cases} X' = AX \\ X(0) \text{ 已指定} \end{cases}$$

的解是 $X(t)=e^{At}X(0)$．

证明 从公式 $X=e^{At}W$，$W=X(0)$，我们立即有

$$\begin{cases}X' = Ae^{At}W = AX \\ X(0) = e^{A0}W = W\end{cases}$$

■

8.11.3 对角阵和可对角化阵

为了在实际运用中使用上面的结果，有必要采用一种有效的方式计算矩阵的指数. 我们从 A 是对角阵的情况开始. 若 $A=\text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$，则容易验证 $A^k=\text{diag}(\lambda_1^k, \lambda_2^k, \cdots, \lambda_n^k)$. 因此，对这样的 A，

600

$$\begin{aligned}e^{At} &= \sum_{k=0}^{\infty}\frac{t^k}{k!}\text{diag}(\lambda_1^k,\lambda_2^k,\cdots,\lambda_n^k) \\ &= \text{diag}(e^{\lambda_1 t},e^{\lambda_2 t},\cdots,e^{\lambda_n t})\end{aligned} \tag{11}$$

在此特殊情况中，微分方程 $X'=AX$ 的解有分量

$$x_i(t) = e^{\lambda_i t}x_i(0) \qquad (1\leqslant i\leqslant n)$$

刚才给出的分析可同时延伸到 A 不是对角阵而是可对角化的情况. 可对角化意指 A 相似于对角阵，换言之，对某个对角阵 Λ 和某个非奇异阵 P，有 $P^{-1}AP=\Lambda$. 若这个等式成立，则如(5)式所示，变量的改变 $X=PY$ 把微分方程 $X'=AX$ 改变成 $Y'=\Lambda Y$. 初始条件 $X(0)$ 变成 $Y(0)=P^{-1}X(0)$，而解是

$$X = PY = P(e^{\Lambda t}P^{-1}X(0)) = P\text{diag}(e^{\lambda_1 t},e^{\lambda_2 t},\cdots,e^{\lambda_n t})P^{-1}X(0) \tag{12}$$

8.11.4 若尔当块

我们直到现在才讨论 A 是不可对角化的情况. 这意味着 $\mathbb{C}^n$ 没有 A 的特征向量组成的基. 这种情况的两个简单的例子由下式给出

$$J(\lambda,2) = \begin{bmatrix}\lambda & 1 \\ 0 & \lambda\end{bmatrix} \qquad J(\lambda,3) = \begin{bmatrix}\lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda\end{bmatrix}$$

我们详细地讨论 $J(\lambda, 3)$. 因为这个矩阵是上三角的，所以它的对角元是其特征值. 因此 $J(\lambda, 3)$ 有单个特征值 λ 并且如果写出等式 $J(\lambda, 3)X=\lambda X$，则我们有

$$\begin{cases}\lambda x_1 + x_2 = \lambda x_1 \\ \lambda x_2 + x_3 = \lambda x_2 \\ \qquad \lambda x_3 = \lambda x_3\end{cases}$$

显然推出 $x_2=x_3=0$，所以仅有形式为 $X=(\beta, 0, 0)^T$ 的解. 此解构成一个一维空间. 换言之，$J(\lambda, 3)$ 的特征向量只张成 $\mathbb{R}^3$ 或 $\mathbb{C}^3$ 中的一个一维子空间. 对每个 $k\geqslant 2$，存在形式为 $J(\lambda, k)$ 的矩阵，它们都具有我们刚才对 $J(\lambda, 3)$ 所说的相同的性质. 这些矩阵称为**若尔当块**. 这一主题的基本定理如下.

601

定理 4(若尔当块定理) 每个方阵相似于一个在它的对角线上有若尔当块的块对角阵.

定理中提到的特殊形式称为给定矩阵的**若尔当标准形**. 下面是几个若尔当标准形的 3×3 矩阵的例子：

$$\begin{bmatrix}7&0&0\\0&5&0\\0&0&3\end{bmatrix}\quad\begin{bmatrix}5&1&0\\0&5&0\\0&0&3\end{bmatrix}\quad\begin{bmatrix}5&0&0\\0&5&1\\0&0&5\end{bmatrix}\quad\begin{bmatrix}0&1&0\\0&0&0\\0&0&0\end{bmatrix}\quad\begin{bmatrix}5&1&0\\0&5&1\\0&0&5\end{bmatrix}$$

其中第一个包含 3 个若尔当块，而第二、第三和第四个包含 2 个若尔当块．第五个本身是 1 个若尔当块．

若尔当块可写成形式

$$J(\lambda,k)=\lambda I_k+H_k \tag{13}$$

其中 I_k 表示 $k\times k$ 单位阵，H_k 表示下列形式的 $k\times k$ 阵：

$$H_k=\begin{bmatrix}0&1&0&\cdots&0&0&0\\0&0&1&\cdots&0&0&0\\0&0&0&\ddots&0&0&0\\\vdots&\vdots&\vdots&\ddots&\ddots&\vdots&\vdots\\0&0&0&\cdots&0&1&0\\0&0&0&\cdots&0&0&1\\0&0&0&\cdots&0&0&0\end{bmatrix} \tag{14}$$

容易看出用 H_k 乘一个向量的影响．我们有

$$\begin{bmatrix}0&1&0&\cdots&0&0\\0&0&1&\cdots&0&0\\0&0&0&\ddots&0&0\\\vdots&\vdots&\vdots&\ddots&\ddots&\vdots\\0&0&0&\cdots&0&1\\0&0&0&\cdots&0&0\end{bmatrix}\begin{bmatrix}\xi_1\\\xi_2\\\xi_3\\\vdots\\\xi_{k-1}\\\xi_k\end{bmatrix}=\begin{bmatrix}\xi_2\\\xi_3\\\xi_4\\\vdots\\\xi_k\\0\end{bmatrix} \tag{15}$$

因为每次应用 H_k 乘 V 消去一个分量，所以显然有 $H_k^kV=0$．因此，H_k 是幂零的；当然，$H_k^k=0$．当 A 是一个若尔当块时，这个事实在计算 e^{At} 中是有用的．因为 H_k 的 k 次幂(以及所有后续幂)为 0，所以我们有

$$\begin{aligned}e^{(\lambda I_k+H_k)t}&=e^{t\lambda I_k}e^{H_kt}\\&=\sum_{j=0}^{\infty}\frac{(\lambda tI_k)^j}{j!}\sum_{j=0}^{\infty}\frac{(tH_k)^j}{j!}\\&=e^{\lambda t}\left[I_k+tH_k+\frac{t^2}{2!}H_k^2+\cdots+\frac{t^{k-1}}{(k-1)!}H_k^{k-1}\right]\end{aligned} \tag{16}$$

注意在前面的计算中我们使用了公式 $e^{A+B}=e^Ae^B$．(见习题 8.11.13.)

602

现在我们求解 A 是若尔当块时的微分方程 $X'=AX$，例如 $A=\lambda I_k+H_k$．应用定理 3 和(16)式得到解；它就是

$$X(t)=e^{At}X(0)=e^{\lambda t}\left[I_k+tH_k+\frac{t^2}{2!}H_k+\cdots+\frac{t^{k-1}}{(k-1)!}H_k^{k-1}\right]X(0) \tag{17}$$

例 2 求解初值问题 $X'=AX$，其中

$$A = \begin{bmatrix} 3 & 1 & 0 & 0 \\ 0 & 3 & 1 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 3 \end{bmatrix} \qquad X(0) = \begin{bmatrix} 7 \\ 5 \\ 3 \\ 9 \end{bmatrix}$$

解 矩阵 A 具有形式 $3I_4 + H_4$. 由(17)式给出的解是

$$X(t) = e^{3t}(I + tH_4 + \frac{1}{2}t^2 H_4^2 + \frac{1}{6}t^3 H_4^3)X(0)$$

$$= e^{3t}\begin{bmatrix} 7 \\ 5 \\ 3 \\ 9 \end{bmatrix} + te^{3t}\begin{bmatrix} 5 \\ 3 \\ 9 \\ 0 \end{bmatrix} + \frac{1}{2}t^2 e^{3t}\begin{bmatrix} 3 \\ 9 \\ 0 \\ 0 \end{bmatrix} + \frac{1}{6}t^3 e^{3t}\begin{bmatrix} 9 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} 7 + 5t + 1.5t^2 + 1.5t^3 \\ 5 + 3t + 4.5t^2 \\ 3 + 9t \\ 9 \end{bmatrix} e^{3t}$$

■

8.11.5 完全一般性解

现在可以描述微分方程组 $X' = AX$ 完全一般性的解. 从 A 的若尔当标准形和把它变为若尔当标准形的相似变换出发. 假如,

$$P^{-1}AP = C$$

其中 C 是 A 的若尔当标准形. 我们知道变量的改变 $X = PY$ 将导致微分方程 $Y' = CY$. (关于这一点见(5)式.) 微分方程 $Y' = CY$ 可分成解开的块. 为此, 考虑下面的例子

603

$$C = \begin{bmatrix} 5 & 1 & 0 & 0 & 0 \\ 0 & 5 & 1 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 7 & 1 \\ 0 & 0 & 0 & 0 & 7 \end{bmatrix}$$

微分方程是

$$\begin{cases} y_1' = 5y_1 + y_2 \\ y_2' = 5y_2 + y_3 \\ y_3' = 5y_3 \\ y_4' = 7y_4 + y_5 \\ y_5' = 7y_5 \end{cases}$$

显然, 第一组 3 个方程可通过它本身求解, 而后面一对也可由它本身求解. 一般情况是十分类似的. 我们看到 C 中的每个若尔当块产生一组微分方程, 它是从保留的方程中解开的. 对每个若尔当块, 可以得到像(17)式中那样的一块解.

例 3 利用上面的矩阵 C, 解具有初始条件 $Y(0) = (3, 2, 8, 4, 1)^T$ 的初值问题 $Y' = CY$.

解 两个解开的系统是

$$\begin{bmatrix} y_1' \\ y_2' \\ y_3' \end{bmatrix} = \begin{bmatrix} 5 & 1 & 0 \\ 0 & 5 & 1 \\ 0 & 0 & 5 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \qquad \text{初值}\begin{bmatrix} 3 \\ 2 \\ 8 \end{bmatrix}$$

$$\begin{bmatrix} y_4' \\ y_5' \end{bmatrix} = \begin{bmatrix} 7 & 1 \\ 0 & 7 \end{bmatrix} \begin{bmatrix} y_4 \\ y_5 \end{bmatrix} \qquad \text{初值} \begin{bmatrix} 4 \\ 1 \end{bmatrix}$$

利用例 2 中说明的方法，我们得到这些解：

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = e^{5t}\left(I + tH_3 + \frac{1}{2}t^2 H_3^2\right)\begin{bmatrix} 3 \\ 2 \\ 8 \end{bmatrix} = e^{5t}\begin{bmatrix} 3 \\ 2 \\ 8 \end{bmatrix} + te^{5t}\begin{bmatrix} 2 \\ 8 \\ 0 \end{bmatrix} + \frac{1}{2}t^2 e^{5t}\begin{bmatrix} 8 \\ 0 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} 3 + 2t + 4t^2 \\ 2 + 8t \\ 8 \end{bmatrix} e^{5t} \begin{bmatrix} y_4 \\ y_5 \end{bmatrix}$$

$$= e^{7t}(I + tH_2)\begin{bmatrix} 4 \\ 1 \end{bmatrix} = e^{7t}\begin{bmatrix} 4 \\ 1 \end{bmatrix} + te^{7t}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = [4 + t1]e^{7t}$$

■

在线性微分方程 $X' = AX$ 的理论中，指数 e^{At} 称为方程的**基本矩阵**. 我们已经看到它是求解相应的微分方程初值问题的关键. 若我们持有 A 的若尔当标准形 C，并知道相似变换

$$P^{-1}AP = C$$

则可以用 $X = PY$ 改变变量，解方程 $Y' = CY$，再回到 X，最终得到

$$X = PY = Pe^{Ct}Y(0) = Pe^{Ct}P^{-1}X(0) \tag{18}$$

604

另一方面，我们知道解的另一种形式 $X = e^{At}X(0)$. 把它与(18)式比较，得出

$$e^{At} = Pe^{Ct}P^{-1}$$

这是一种计算基本矩阵的方法.

8.11.6 非齐次问题

已经建立的原则现在可应用于**非齐次问题**

$$X' = AX + W \tag{19}$$

其中 W 可以是 t 的一个向量函数. 首先我们考虑 A 是可对角化的情况. 相似变换 $P^{-1}AP = \Lambda$ 产生一个对角阵 Λ. 变量替换 $X = PY$ 变换(19)式为

$$Y' = \Lambda Y + P^{-1}W \tag{20}$$

这是一组完全分离的 n 个方程，其中典型的一个具有形式

$$\eta'(t) = \lambda\eta(t) + g(t) \tag{21}$$

这个方程的解是

$$\eta(t) = e^{\lambda t}\left[\eta(0) + \int_0^t e^{-\lambda s}g(s)\,ds\right] \tag{22}$$

例 4 设

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \qquad W = \begin{bmatrix} t^2 \\ t \\ \sin t \end{bmatrix} \qquad X(0) = \begin{bmatrix} 5 \\ 7 \\ 9 \end{bmatrix}$$

求解方程 $X' = AX + W$.

解 这个方程组由下列单独的解开的方程组成

$$\begin{cases} x_1' = t^2 \\ x_2' = x_2 + t \\ x_3' = 2x_3 + \sin t \end{cases}$$

由(22)式得到的解是

$$x_1(t)=5+\int_0^t s^2\mathrm{d}s=5+\frac{1}{3}t^3$$

[605]

$$x_2(t)=\mathrm{e}^t\left[7+\int_0^t \mathrm{e}^{-s}s\,\mathrm{d}s\right]=8\mathrm{e}^t-t-1$$

$$x_3(t)=\mathrm{e}^{2t}\left[9+\int_0^t \mathrm{e}^{-2s}\sin s\,\mathrm{d}s\right]=\frac{46}{5}\mathrm{e}^{2t}-\frac{2}{5}\sin t-\frac{1}{5}\cos t$$ ■

如果方程 $X'=AX+W$ 中的矩阵 A 不可对角化，则我们可使用若尔当标准形和一个变量替换把问题分开为解开的子系统. 因此，用一个单独的若尔当块来说明这个过程就足够了.

例 5 设

$$A=\begin{bmatrix}5&1&0\\0&5&1\\0&0&5\end{bmatrix}\qquad W=\begin{bmatrix}t^2\\t\\\sin t\end{bmatrix}\qquad X(0)=\begin{bmatrix}5\\7\\9\end{bmatrix}$$

求解方程 $X'=AX+W$.

解 单独的方程是

$$\begin{cases}x_1'=5x_1+x_2+t^2\\x_2'=5x_2+x_3+t\\x_3'=5x_3+\sin t\end{cases}$$

这里推荐的过程是从末端开始以相反的次序解方程. 这就得到

$$\begin{aligned}x_3(t)&=\mathrm{e}^{5t}\left\{9+\int_0^t \mathrm{e}^{-5s}\sin s\,\mathrm{d}s\right\}\\&=\left(9+\frac{1}{26}\right)\mathrm{e}^{5t}-\frac{5}{26}\sin t-\frac{1}{26}\cos t\\x_2(t)&=\mathrm{e}^{5t}\left\{7+\int_0^t \mathrm{e}^{-5s}[x_3(s)+s]\mathrm{d}s\right\}\\&=\left(7+\frac{1}{25}-\frac{10}{26^2}\right)\mathrm{e}^{5t}+\frac{1}{2}\left(9+\frac{1}{26}\right)t\mathrm{e}^{5t}\\&\quad+\frac{2}{26^2}(12\sin t+t\cos t)-\frac{1}{5}t-\frac{1}{25}x_1(t)\\x_1(t)&=\mathrm{e}^{5t}\left\{5+\int_0^t \mathrm{e}^{-5s}[x_2(s)+s^2]\mathrm{d}s\right\}\\&=\left(5+\frac{74}{26^3}\right)\mathrm{e}^{5t}+\left(7+\frac{1}{25}-\frac{10}{26^2}\right)t\mathrm{e}^{5t}+\frac{1}{2}\left(9+\frac{1}{26}\right)t^2\mathrm{e}^{5t}\\&\quad-\frac{2}{26^3}(55\sin t+37\cos t)-\frac{1}{5}t^2-\frac{1}{25}t\end{aligned}$$ ■

在计算矩阵指数 e^A 的工作中必须充分认识到可能的隐患. 我们对读者推荐 Moler and Van Loan[1978]的论文. 他们的研究指出下列 4 步过程在大多数情况下是适用的：

[606]

1. 设 j 是使 $\|A\|/2^j\leqslant 1/2$ 的第一个正整数.
2. 若 ϵ 是给出的相对误差容限，选择 p 是使 $2^{p-3}(p+1)\geqslant 1/\epsilon$ 的第一个正整数.

3. 根据截断泰勒级数 $e^z=\sum_{k=0}^{p} z^k/k!$ 计算 $e^{A/2^j}$.

4. 把第 3 步计算的 $e^{A/2^j}$ 平方 j 次得到 $e^A=(e^{A/2^j})^{2^j}$.

从第 4 步得到的 e^A 的计算值是 e^{A+E}，其中 E 是一个满足 $\|E\|/\|A\|\leqslant\epsilon$ 的矩阵. 这个结论引用的是参考文献中的推论 1.

习题 8.11

1. 设

$$A=\begin{bmatrix}1&0&3\\-1&1&-1\\3&0&1\end{bmatrix}$$

求解方程 $X'=AX$ 的通解.

2. (续)求上题中的方程服从初始条件 $X_0=(-1,\ 4,\ 7)^T$ 的解.

3. 求下列方程组的通解

$$\begin{cases}x_1'=3x_1-5x_2\\x_2'=2x_1+x_2\end{cases}$$

4. 对任意 $n\times n$ 的矩阵 A，证明：e^A 是非奇异的. 注意：没有假设对所有 $n\times n$ 矩阵 A 和 B 有 $e^Ae^B=e^{A+B}$. (见下题.)

5. (续)用例子说明等式 $e^Ae^B=e^{A+B}$ 不总是成立的. 提示：考虑 e^{At}，e^{Bt} 和 $e^{(A+B)t}$.

6. 证明：若 $A=P^{-1}BP$，则 $e^A=P^{-1}e^BP$.

7. 详细说明如何利用(8)式的估计以及完整的论证建立(6)式收敛性.

8. 对 k 用归纳法证明：若 V_1，V_2，…，V_k 是矩阵对应于 k 个不同的特征值的特征向量，则 $\{V_1,\ V_2,\ \cdots,\ V_k\}$ 线性无关.

9. 通过对 e^{At} 的级数的微分证明(10)式.

10. 利用上面习题 8.11.6 中的结论，不作变量替换直接建立(12)式中的结果.

11. 设 B 是 $n\times n$ 矩阵，并设 V_1，V_2，…，V_k 是 $\mathbb{R}^n$ 中的向量使得 $V_1\neq0$，$BV_1=0$，$BV_2=V_1$，…，$BV_k=V_{k-1}$. 利用对 k 的归纳法证明 $\{V_1,\ V_2,\ \cdots,\ V_k\}$ 线性无关. k 能够为多大?

12. 对一个二阶矩阵，求出使它的若尔当标准形只含一个若尔当块的确切条件.

13. 证明 $e^{A+B}=e^Ae^B$ 当且仅当 $AB=BA$. (见上面习题 8.11.5.)

14. 证明：若 A 和 B 用同样的相似变换对角化(即，PAP^{-1} 和 PBP^{-1} 都是对角阵)，则 $AB=BA$.

15. 证明线性方程组

$$X'=AX+V(t)\qquad X(0)=W$$

的解为

$$X(t)=e^{At}W+e^{At}\int_0^t e^{-As}V(s)\mathrm{d}s$$

说明这里出现的不定积分代表什么.

16. 当 X 在不是 0 的点 t_0 上给定时，初值问题 $X'=AX$ 的解是什么?

17. 说明方程组

$$X'=AX\quad,\quad A=\begin{bmatrix}-1&6\\1&-2\end{bmatrix}$$

607

的基础矩阵是

$$\frac{1}{5}\begin{bmatrix} 2e^{-4t}+3e^{t} & -6e^{-4t}+6e^{t} \\ -e^{-4t}+e^{t} & 3e^{-4t}+2e^{t} \end{bmatrix}$$

18. 证明：基本矩阵中的第 j 列是初值问题 $X'=AX$，$X(0)=U_j$ 的解，其中 U_j 是第 j 个标准单位向量.

19. 对 e^{A} 的逆作一个猜想，然后证明你是对的. 注意：一般说来 $e^{A+B}\neq e^{A}e^{B}$.

20. 设

$$A=\begin{bmatrix} 5 & 4 & 3 \\ -1 & 0 & -3 \\ 1 & -2 & 1 \end{bmatrix}, \quad X(0)=\begin{bmatrix} 1 \\ 5 \\ 3 \end{bmatrix}$$

求解 $X'=AX$.

21. 完成例 4.

22. 证明：一个不可对角化的矩阵是可对角化矩阵序列的极限.

23. (续)考虑线性方程组 $X'=AX$，$X(0)=V$，其中 A 不是可对角化的. 若 B 是一个接近于 A 的可对角化阵，则关于 $Y'=BY$，$Y(0)=V$ 的解可说些什么？

24. 证明：$\det e^{A}=e^{\operatorname{tr}(A)}$，其中 A 是任意的 $n\times n$ 矩阵而 $\operatorname{tr}(A)$是 A 的迹——即 A 中对角元素之和.

25. 对 $n\times n$ 阵 A，设 $|A|=\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|$. 证明

$$|e^{A}|\leqslant n-1+e^{|A|}$$

26. 尽管一般说来 $e^{A+B}\neq e^{A}e^{B}$，但是可证明 $e^{A+B}=\lim_{k\to\infty}[e^{A/k}e^{B/k}]^{k}$.

8.12 刚性方程

常微分方程组中的**刚性**指的是向量解中分量的时间尺度相差悬殊. 通常某些相当好的数值方法对刚性方程用起来很差. 当数值解中的稳定性只能以非常小的步长才能达到时发生这种情况.

刚性方程产生在一些应用中. 例如，在空间飞行器的控制中，预期的飞行路径是十分光滑的，但是当发现计划的飞行路径出现任何偏离时在过程中可以作出非常迅速的校正. 这种问题的另一个来源是化学过程的监控，因为物理的和化学的变化可能存在时间尺度上的巨大差异. 在电子电路理论中，因为微秒级的瞬间变化可能强加在大致平稳的电路上，所以产生刚性问题.

608

8.12.1 欧拉方法

可对一个简单的模型问题用欧拉方法来说明数值上的难点. 欧拉方法是一阶泰勒方法，对初值问题

$$\begin{cases} x'=f(t,x) \\ x(t_0)=x_0 \end{cases} \tag{1}$$

它利用下式进行：

$$x_{n+1}=x_n+hf(t_n,x_n) \quad (n\geqslant 0) \tag{2}$$

现在考虑对简单的试验问题

$$\begin{cases}x' = \lambda x \\ x(0) = 1\end{cases} \tag{3}$$

利用欧拉方法的结果. 用欧拉方法产生数值解

$$\begin{cases}x_0 = 1 \\ x_{n+1} = x_n + h\lambda x_n = (1 + h\lambda)x_n\end{cases} \tag{4}$$

因此，在第 n 步，近似解为

$$x_n = (1 + h\lambda)^n \tag{5}$$

另一方面，方程(3)的实际解是

$$x(t) = e^{\lambda t} \tag{6}$$

若 $\lambda<0$，则(6)式中的解是指数衰减的. 当 t 达到无穷时它趋于 0 的稳定状态. (5)式中的数值解趋于 0 当且仅当 $|1+h\lambda|<1$. 这迫使我们选择 $h>0$ 以使得 $1+h\lambda>-1$. 因为 $\lambda<0$，所以必须强加条件 $h<-2/\lambda$.

例如，若 $\lambda=-20$，虽然我们尝试跟踪的解在初始时刻 $t=0$，$x=1$ 以后不久极为平坦(特别在 0 点)，但是必须取 $h<0.1$. 我们注意到 $x(t)=e^{-20t}\leqslant 2.1\times 10^{-9}$，$t\geqslant 1$. 因此，在区域中数值解必须用小步长继续进行，而解的性质指出可以采用大步长. 这个现象是刚性的一个方面. 像 e^{-20t} 那样的函数几乎直接(即以非常大的负斜率)衰减到 0 称为**瞬变现象**，因为它的物理作用具有短暂的持续时间. 我们期望瞬变函数的数值轨迹需要小步长，直到瞬变作用变得忽略不计为止；之后，一个好的数值方法应该允许大步长. 但欧拉方法不满足这个要求. [609]

8.12.2 修正的欧拉方法

相反，隐式的欧拉方法可以符合刚才提到的准则. 隐式的欧拉方法由下式定义

$$x_{n+1} = x_n + hf(t_{n+1}, x_{n+1}) \quad (n \geqslant 0) \tag{7}$$

当这个方法应用于(3)式中的试验问题时，我们得到

$$\begin{cases}x_0 = 1 \\ x_{n+1} = x_n + h\lambda x_{n+1}\end{cases} \tag{8}$$

这给出

$$x_{n+1} = (1 - h\lambda)^{-1}x_n \tag{9}$$

因此，在第 n 步，我们有

$$x_n = (1 - h\lambda)^{-n} \tag{10}$$

对负的 λ，数值解应该酷似实际解的要求变成

$$|1 - h\lambda|^{-1} < 1 \tag{11}$$

这对一切(正的)步长 h 显然成立.

8.12.3 微分方程组

类似的考虑应用于微分方程组. 我们再次求助于这个原则，就是说一个好的数值方法应该对简单的线性试验情况执行得很好. (记得在 8.5 节中使用这个原则推出：稳定性和相容性是公认的线性多步法具有的基本特性.) 下面是一个简单的试验案例，它是一个包含两个微分方程的方程组：

$$\begin{cases} x' = \alpha x + \beta y & x(0) = 2 \\ y' = \beta x + \alpha y & y(0) = 0 \end{cases} \tag{12}$$

其解是

610

$$\begin{cases} x(t) = e^{(\alpha+\beta)t} + e^{(\alpha-\beta)t} \\ y(t) = e^{(\alpha+\beta)t} - e^{(\alpha-\beta)t} \end{cases} \tag{13}$$

若用欧拉方法计算方程(12)的数值解，则前进求解的公式是

$$\begin{cases} x_{n+1} = x_n + h(\alpha x_n + \beta y_n) & x_0 = 2 \\ y_{n+1} = y_n + h(\beta x_n + \alpha y_n) & y_0 = 0 \end{cases} \tag{14}$$

这些差分方程的解是

$$\begin{cases} x_n = (1 + \alpha h + \beta h)^n + (1 + \alpha h - \beta h)^n \\ y_n = (1 + \alpha h + \beta h)^n - (1 + \alpha h - \beta h)^n \end{cases} \tag{15}$$

我们感兴趣的情况由 $\alpha<\beta<0$ 所确定. 用这个假设，方程(13)的解按指数衰减到0. 为了使方程(15)中的数值解酷似这个性态，我们要求

$$|1 + \alpha h + \beta h| < 1 \quad |1 + \alpha h - \beta h| < 1 \tag{16}$$

因为由 $\alpha<\beta<0$，所以(16)中的不等式等价于单个条件

$$0 < h < -2/(\alpha+\beta)$$

为看看能发生什么情况，假设 $\alpha=-20$ 且 $\beta=-19$. 则我们的解是 e^{-39t} 和 e^{-t} 的组合. 其中第1个是瞬变函数，在一个短暂的时间区间以后它与 e^{-t} 相比较是可忽略不计的. 但是瞬变将控制遍及整个数值解的允许的步长.

8.12.4 一般的线性多步法

我们考察8.5节中一般的线性多步法. 为了对单个试验方程(3)满意地执行这个方法，看看它应该有什么附加的性质. 一般的方法有下列形式

$$a_k x_n + a_{k-1} x_{n-1} + \cdots + a_0 x_{n-k} = h[b_k f_n + b_{k-1} f_{n-1} + \cdots + b_0 f_{n-k}] \tag{17}$$

当这个式子应用于试验方程时，我们得到

$$a_k x_n + a_{k-1} x_{n-1} + \cdots + a_0 x_{n-k} = h\lambda(b_k x_n + b_{k-1} x_{n-1} + \cdots + b_0 x_{n-k}) \tag{18}$$

因此，数值解将求解齐次线性差分方程

$$(a_k - h\lambda b_k) x_n + (a_{k-1} - h\lambda b_{k-1}) x_{n-1} + \cdots + (a_0 - h\lambda b_0) x_{n-k} = 0 \tag{19}$$

这个方程的解将是类型为 $x_n = r^n$ 的某些基本解的组合，这里 r 是多项式

611

$$\phi(z) = (a_k - h\lambda b_k) z^k + (a_{k-1} - h\lambda b_{k-1}) z^{k-1} + \cdots + (a_0 - h\lambda b_0) \tag{20}$$

的一个根. 注意(20)式中定义的多项式具有形式 $\phi = p - \lambda h q$，其中 p 和 q 是8.5节中所用的多项式，即

$$p(z) = a_k z^k + a_{k-1} z^{k-1} + \cdots + a_1 z + a_0 \tag{21}$$

$$q(z) = b_k z^k + b_{k-1} z^{k-1} + \cdots + b_1 z + b_0 \tag{22}$$

8.12.5 *A* 稳定性

若试验问题(3)中的 $\lambda<0$，则解是指数衰减的. 为使从(17)式中的多步法得到的数值解反映这个性态，(20)式中的多项式 ϕ 的所有根必须位于圆盘 $|z|<1$ 中. 对 λ 的复数值 $\lambda=\mu+$

iv，试验问题的解是

$$x(t)=e^{\lambda t}=e^{\mu t}e^{i\nu t}=e^{\mu t}(\cos\nu t+i\sin\nu t)$$

在此情况下指数衰减对应于 $\mu<0$. 为了使多步法顺利执行这个问题，只要 $h>0$ 且 $\mathrm{Re}(\lambda)<0$，我们希望 ϕ 的根出现在单位圆盘的内部. 这个性质称为 **A 稳定性.**

如(11)式中所示，(7)式中的隐式欧拉方法是 A 稳定的. 由

$$x_n-x_{n-1}=\frac{1}{2}h[f_n+f_{n-1}] \tag{23}$$

定义的隐式梯形方法也是 A 稳定的，因为多项式 ϕ 是

$$\phi(z)=z-1-\lambda h\left(\frac{1}{2}z+\frac{1}{2}\right)$$

并且它的根是 $z=(2+\lambda h)/(2-\lambda h)$；容易看出当 $h>0$ 且 $\mathrm{Re}(\lambda)<0$ 时，这个根是在单位圆盘内部.

Dahlquist[1963]的一个重要定理宣称一个 A 稳定的线性多步法必是一个隐式方法，并且它的阶不超过 2. 这个结果对 A 稳定方法加上了一个严格的限制. 隐式梯形法则通常用于刚性方程，因为它在所有的 A 稳定多步法中具有最小的截断误差.

8.12.6 绝对稳定性区域

每个多步法都有一个**绝对稳定性区域**. 这是复数 ω 的集合，它使得 $p-\omega q$ 的根位于单位圆盘的内部. 从前面的讨论可得：若 λh 位于绝对稳定性区域中，则对试验问题 $x'=\lambda x$ 的方法就能顺利工作. 我们也注意到一个方法是 A 稳定的当且仅当它的绝对稳定性区域包含左半平面.

例 1 欧拉方法的绝对稳定性区域是什么?

解 欧拉方法由下式定义

612

$$x_n-x_{n-1}=hf_{n-1}$$

因此，$p(z)=z-1$，$q(z)=1$，$\phi(z)=z-1-\omega$，其中 $\omega=\lambda h$. ϕ 的根是 $z=1+\omega$. 为使 ϕ 的所有根在单位圆盘的内部，我们需要 $|1+\omega|<1$. 这是一个复平面上中心在 -1 的半径为 1 的圆盘. ■

因为 A 稳定性是如此严格的限制的，所以不具有这种性质的方法是决不用于刚性方程的. 当我们使用方法时，希望 $\omega=\lambda h$ 位于方法的绝对稳定性区域中. λ 附属于求解的微分方程. 对单个线性方程，λ 是方程中 x 的系数. 对单个非线性方程，λ 可能是用线性方程局部地逼近方程产生的一个常数. 对线性方程组 $X'=AX$，λ 可能是 A 的一个特征值. 因此，h 倍 A 的每个特征值应该位于所用方法的绝对稳定性区域中. 对非线性方程组，我们可以试图用线性方程组局部逼近给定的方程组，并应用前面的准则. 这个想法通常是不可能实施的，因此，在困难的刚性问题情况中最合适的安全策略可能是回复到梯形法则. 初值问题的某些最流行的代码努力查出步进求解过程中的刚性，并且当出现刚性时采取合适的防御措施. 有关的细节建议读者查阅 Gear[1971]以及 Shampine and Gordon[1975]的文献. 一般说来，高阶方法有较小的绝对稳定性区域，并且当存在刚性时，这些自适应代码会转换到具有更有利的绝对稳定性区域的低阶方法.（见 Byrne and Hindmarsh[1987]或 Shampine and Gear[1979].）

8.12.7 非线性方程

为考察前面的研究与非线性方程组的相关性，我们转向一个典型的假定是自控的方程组：

$$X' = F(X) \tag{24}$$

这里 X 是 t 的向量函数，它有 n 个分量函数 x_1，x_2，…，x_n. 右边有一个函数 F：$\mathbb{R}^n \to \mathbb{R}^n$. 这个函数有分量函数 f_i. F 的**雅可比矩阵**是矩阵 $J=(J_{ij})$，它定义为 $J_{ij} = \partial f_i / \partial x_j$. 在点 X_0，方程(24)的线性化形式是

$$X' = F(X_0) + J(X_0)(X - X_0) \tag{25}$$

其中 $J(X_0)$表示 J 在 X_0 的值. (25)式是一个线性微分方程，矩阵 $J(X_0)$的特征值在理论上引

613

人注目. 若方程(24)在 X_0 附近是刚性的话，则这些特征值满足不等式

$$\mathrm{Re}(\lambda_1) \leqslant \mathrm{Re}(\lambda_2) \leqslant \cdots \leqslant \mathrm{Re}(\lambda_n) < 0$$

并且 $\mathrm{Re}(\lambda_1)$比 $\mathrm{Re}(\lambda_n)$小得多.

若多步法对这样的问题执行顺利的话，则 $h\lambda_i$ 应该位于这个方法的绝对稳定性区域内. 若 λ_i 已知，则可利用这个信息适当地选择 h.

习题 8.12

1. 求由(7)式定义的隐式欧拉方法的绝对稳定性区域.
2. 确定问题 $x''=(57+\sin t)x$ 是否为刚性的.
3. 确定问题
$$\begin{cases} x'' = -20x' - 19x \\ x(0) = 2 \quad x'(0) = -20 \end{cases}$$
是否为刚性的.
4. 考虑多步法
$$x_n + \alpha x_{n-1} - (1+\alpha)x_{n-2} = \frac{1}{2}h[-\alpha f_n + (4+3\alpha)f_{n-1}]$$
确定 α 使得方法是稳定的，相容的，收敛的，A 稳定的和二阶的.（分别见 8.4 节和 8.5 节.）
5. 验证方程(15)提供方程(14)的解.
6. 求(23)式隐式梯形法则的绝对稳定性区域.
7. 当 $\alpha<\beta<0$ 时，证明两个不等式关系(16)等价于 $0<h<-2/(\alpha+\beta)$.

计算机习题 8.12

1. 求解具有初始条件 $x(0)=1$，$x'(0)=-\sqrt{57}$的习题 8.12.2 中的微分方程(**Riccati 变换**是 $y=x'/x$. 它可以用来得到一对等价的方程，$y'=57+\sin(t)-y^2$，$x'=xy$，且 $y(0)=-\sqrt{57}$，$x(0)=1$，使得能够应用共享软件.)
2. 对方程 $x'=\lambda x$，$\lambda<0$ 检验**隐式中点方法**
$$x_n - x_{n-1} = hf\left(t_{n-1} + \frac{1}{2}h, \frac{1}{2}(x_n + x_{n-1})\right)$$
确定它的性能对刚性方程是否令人满意.
3. 用刚性方程组
$$\begin{cases} x_1' = -1\,000x_1 + x_2 \quad x_1(0) = 1 \\ x_2' = 999x_1 - 2x_2 \quad x_2(0) = 0 \end{cases}$$
描述某个化学反应. 证明 x_1 迅速地衰减而 x_2 缓慢地衰减. 因为需要一个非常小的步长，所以这样的一个

614

方程组的求解是困难的.

第9章　偏微分方程数值解

9.0　概述

本章介绍偏微分方程的数值解问题．在这个领域中出现的数值计算容易过度地耗费最大和最快的计算机资源．因为在求解偏微分方程中通常包含巨大的计算工作量，数值分析这一分支是当前研究十分活跃的一个分支．当我们考虑几个代表性的问题及其有效的求解过程时，这些问题的存储量和运行时间(即使在超级计算机上)为什么如此巨大就变得清晰了．

9.1　抛物型方程：显式方法

9.1.1　热传导方程

我们从**抛物型**的一个代表性的偏微分方程——**热传导方程**，也称为**扩散方程**开始讨论．如果适当地选取物理量的单位，则热传导方程具有下列形式：

$$u_{xx}+u_{yy}+u_{zz}=u_t \tag{1}$$

在这个方程中，u 是一个 x，y，z 和 t 的函数．记号 u_t 表示 $\partial u/\partial t$，u_{xx} 表示 $\partial^2 u/\partial x^2$ 等等．方程(1)决定时间 t 及三维体中位置(x, y, z)上的温度 u．我们用 $u(x, y, z, t)$表示 u 在点(x, y, z, t)上的值．因此，u 是 4 个实变量的实值函数．

615

正如在常微分方程的理论中那样，一个正确提出的物理问题决不是单独由一个偏微分方程组成的；为了确定问题的唯一解，必须存在足够数目的附加的边界条件．下面我们用一维形式的热传导方程(1)以及附属的条件作为一个模型问题

$$\begin{cases} u_{xx}=u_t & (t\geqslant 0, 0\leqslant x\leqslant 1)\\ u(x,0)=g(x) & (0\leqslant x\leqslant 1)\\ u(0,t)=a(t) & (t\geqslant 0)\\ u(1,t)=b(t) & (t\geqslant 0)\end{cases} \tag{2}$$

方程组(2)模拟长度为 1 的一根杆中的温度分布；其端点分别保持温度 $a(t)$和 $b(t)$．(见图 9-1.)假定函数 g，a 和 b 是给定的．而且初始温度剖面由函数 g 规定．函数 u 是(x, t)的函数，它不依赖于 y 和 z．图 9-2 显示了要求的 $u(x, t)$的区域，它是由变量 x 和 t 确定的平面的一个子集．

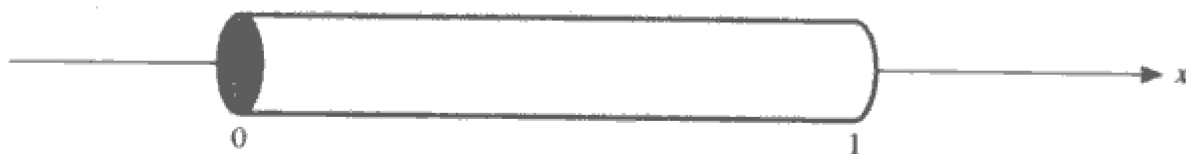


图 9-1　单位长杆

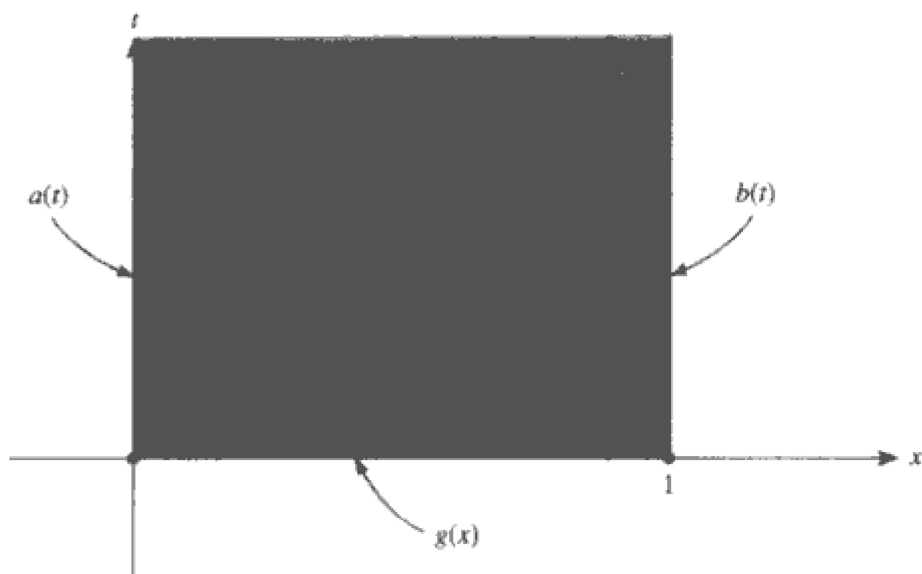


图 9-2 解域

9.1.2 有限差分法

[616] 用于数值求解像(2)那样问题的主要方法之一是**有限差分法**. 它涉及区域的初始**离散化**如下:

$$\begin{cases} t_j = jk & (j \geqslant 0) \\ x_i = ih & (0 \leqslant i \leqslant n+1) \end{cases} \tag{3}$$

变量 t 和 x 有不同的步长，分别用 k 和 h 表示. 因为 x 取遍区间[0, 1]，所以 $h=1/(n+1)$. 我们的目标是在所谓的**网格点**(x_i, t_j)上计算解函数 u 的近似值.

过程的下一步是选择某些简单的公式去逼近微分方程中出现的导数. 可使用的由(3)式、(8)式和(9)式得到的一些熟悉的基本公式是

$$f'(x) \approx \frac{1}{h}[f(x+h) - f(x)] \tag{4}$$

$$f'(x) \approx \frac{1}{2h}[f(x+h) - f(x-h)] \tag{5}$$

$$f''(x) \approx \frac{1}{h^2}[f(x+h) - 2f(x) + f(x-h)] \tag{6}$$

当然，存在许多其他这样的公式，它们提供各种精确度.（见 7.1 节习题.）

现在，我们利用(4)和(6)式，用微分问题(2)的离散化形式代替微分问题. 因为这两个问题有不同的解，所以对离散问题用另外的字母 v:

$$\frac{1}{h^2}[v(x+h,t) - 2v(x,t) + v(x-h,t)] = \frac{1}{k}[v(x,t+k) - v(x,t)] \tag{7}$$

为简化记号，我们在(7)式中设 $x=x_i$ 和 $t=t_j$，则 $v(x_i, t_j)$缩写为 v_{ij}. 结果是

$$\frac{1}{h^2}(v_{i+1,j} - 2v_{ij} + v_{i-1,j}) = \frac{1}{k}(v_{i,j+1} - v_{ij}) \tag{8}$$

当(8)式中的 $j=0$ 时，所有项除 v_{i1} 外都已知. 当然，初始温度分布 g 给出

$$g(x_i) = u(x_i,0) = v_{i0}$$

换言之，我们知道对应于 $t=0$ 的 t 层上的 u（以及 v）的正确值. 所以 v_{i1} 的值可从(8)式计算.

为此，(8)式可写成等价的形式

$$v_{i,j+1}=\frac{k}{h^2}(v_{i+1,j}-2v_{i,j}+v_{i-1,j})+v_{i,j}$$

或者利用缩写 $s=k/h^2$，

$$v_{i,j+1}=sv_{i-1,j}+(1-2s)v_{ij}+sv_{i+1,j} \tag{9}$$

617

利用(9)式，数值解可以在 t 方向一步步前进. 图 9-3 中的略图显示了(9)式中涉及的一组代表性的 4 个网格点. 因为(9)式利用前面的值 $v_{i-1,j}$，v_{ij}，$v_{i+1,j}$ 显式地给出新值 $v_{i,j+1}$，所以基于这个式子的方法称为显式方法.

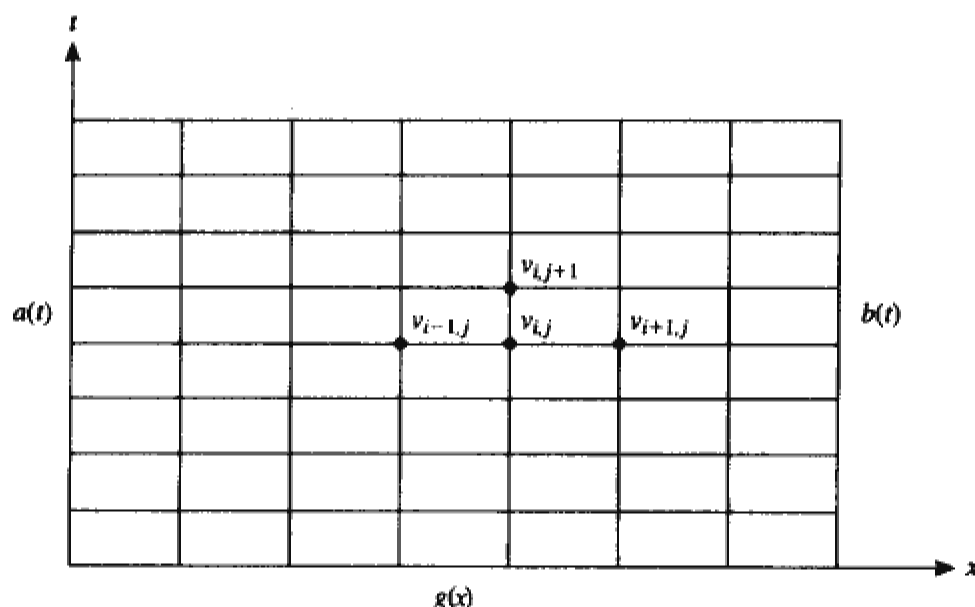


图 9-3 典型的网格

我们强调，即使相继 v_{ij} 的计算完全精确地执行，数值解也不同于偏微分方程的解. 这是因为函数 v 是一个不同问题的解，即微分方程的有限差分模拟的解. 事实上，因为在近似微分公式中使用低精度，所以我们预料到有一个相当大的差别.

9.1.3 算法

执行前面显式方法计算的算法包含两个阶段：初始化和求解. 这里，在[0，1]中网格点 x_i 的个数为 $n+2$，t 变量的步长为 k，t 中计算的步数是 M.

```
input n, k, M
h ← 1/(n+1)
s ← k/h²
w_i ← g(ih)      (0 ≤ i ≤ n+1)
t ← 0
output 0, t, (w_0, w_1, ..., w_{n+1})
for j = 1 to M do
    v_0 ← a(jk)
    v_{n+1} ← b(jk)
    for i = 1 to n do
        v_i ← s w_{i-1} + (1-2s) w_i + s w_{i+1}
    end do
```

618

```
    t←jk
    output j, t, (v0, v1, ..., vn+1)
    (w1, w2, ..., wn)←(v1, v2, ..., vn)
end do
```

我们极力主张读者对算法进行编程并执行计算机习题 9.1.1 中所述的数值实验. 此类实验产生这样的结论：不是所有的步长对(h, k)均令人满意. 下段中的分析将说明为什么这个结论是正确的. 我们将分析 $a(t)=b(t)=0$ 时的特殊情况.

9.1.4 稳定性分析

定义数值过程的(9)式可用矩阵-向量记号说明. 设 $t=jk$ 时刻值向量用 V_j 表示. 因此，

$$V_j = \begin{bmatrix} v_{1j} \\ v_{2j} \\ \vdots \\ v_{nj} \end{bmatrix} \tag{10}$$

(9)式可写成

$$V_{j+1} = AV_j \tag{11}$$

其中 A 是 $n\times n$ 矩阵

$$A = \begin{bmatrix} 1-2s & s & 0 & \cdots & 0 \\ s & 1-2s & s & \cdots & 0 \\ 0 & s & 1-2s & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1-2s \end{bmatrix} \tag{12}$$

注意：因为 $a(t)=b(t)=0$，所以 $v_{0j}=v_{n+1,j}=0$. 因此，在每条水平线上仅出现 n 个未知值.

在这个阶段，论证可用两种方法进行. 第一种论证与物理事实有关，因为杆的端点保持温度为 0，所以当 $t\to\infty$时，杆中的温度将趋于 0. 因此我们必须要求当 $t\to\infty$时，数值解也趋于 0. 因为 $V_{j+1}=AV_j$，所以有

$$V_j = AV_{j-1} = A^2V_{j-2} = \cdots = A^jV_0$$

由 4.6 节定理 5，下列两个条件等价：

1. 对一切向量 V，$\lim\limits_{j\to\infty}A^jV=0$.

619

2. $\rho(A)<1$.

记得 $\rho(A)$是矩阵 A 的谱半径，因此，应该选择参数 $s=k/h^2$ 使得 $\rho(A)<1$.

如果不要求分析温度分布的物理问题，代之分析数值计算中舍入误差的影响也可以得到同样的结论. 假如在某一步(可假定是第 1 步)引进误差. 则代替向量 V_0，我们有它的一个扰动 $\widetilde{V}_0$. 显式方法将产生向量 $\widetilde{V}_j=A^j\widetilde{V}_0$，在第 j 步的误差是

$$V_j - \widetilde{V}_j = A^jV_0 - A^j\widetilde{V}_0 = A^j(V_0 - \widetilde{V}_0)$$

为保证这个误差在 $j\to\infty$时消失，仍然要求 $\rho(A)<1$.

特征值的确定将在后面给出；计算结果得到 A 的特征值是

$$\lambda_j = 1 - 2s(1 - \cos\theta_j) \qquad \theta_j = \frac{j\pi}{n+1} \qquad (1 \leqslant j \leqslant n) \tag{13}$$

为使 $\rho(A)$小于 1，我们要求

$$-1 < 1 - 2s(1 - \cos\theta_j) < 1$$

当且仅当 $s<(1-\cos\theta_j)^{-1}$此式成立. 因为 s 是正的，所以对 s 最大的限制出现在 $\cos\theta_j = -1$ 时. 因为当 $j=n$ 时 θ_j 十分接近于 π，所以必须要求 $s\leqslant 1/2$.

小结：为使前面的显式算法**稳定**，必须假定 $s=k/h^2\leqslant 1/2$.

严格的限制 $k\leqslant h^2/2$ 使得这个方法非常缓慢. 例如，若 $h=0.01$，则 k 的最大允许值是5×10^{-5}. 若想对于 $0<t<10$ 计算解，则时间步数必须为$\frac{1}{2}\times 10^6$ 而网格点数大于 2 千万！故显式方法的漂亮和简洁性伴随无法接受的低效率.

为完成前面的分析，必须证明 A 的特征值为(13)式中所给出的那样. 首先注意(11)式中的 A 可写成

$$A = I - sB$$

其中 B 是 $n\times n$ 矩阵

$$B = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 2 \end{bmatrix} \tag{14}$$

A 的特征值λ_i 和 B 的特征值μ_j 的关系为

$$\lambda_i = 1 - s\mu_i$$

因此，只要确定 B 的特征值就足够了.

620

引理 1(三对角阵特征值和特征向量引理) *设 $x=(\sin\theta, \sin 2\theta, \cdots, \sin n\theta)^{\mathrm{T}}$. 若 $\theta=j\pi/(n+1)$，则 x 是 B 对应于特征值 $2-2\cos\theta$ 的特征向量.*

证明 设 $B_{ij}=2\delta_{ij}-\delta_{i+1,j}-\delta_{i-1,j}$，其中 $B=(B_{ij})$而 δ_{ij} 是克罗内克符号. 因此，若 $2\leqslant i\leqslant n-1$,则

$$\begin{aligned}(Bx)_i &= \sum_{j=1}^{n}(2\delta_{ij} - \delta_{i+1,j} - \delta_{i-1,j})x_j \\ &= 2x_i - x_{i+1} - x_{i-1} \\ &= 2\sin i\theta - [\sin(i+1)\theta + \sin(i-1)\theta] \\ &= 2\sin i\theta - 2\sin i\theta\cos\theta \\ &= (2 - 2\cos\theta)\sin i\theta \\ &= (2 - 2\cos\theta)x_i\end{aligned}$$

这里我们用了标准关系 $\sin(\alpha+\beta)+\sin(\alpha-\beta)=2\sin\alpha\cos\beta$. 若 $i=1$ 或 $i=n$，倘若 $x_0=0$ 和 $x_{n+1}=0$ 则我们可利用相同的计算. 公式 $x_i=\sin i\theta$ 自动地给出 $x_0=0$，且当 $\sin(n+1)\theta=0$ 或对某个整数 j，$(n+1)\theta=j\pi$ 时，等式 $x_{n+1}=0$ 成立. 因此两个特殊情况($i=1$ 及 $i=n$)为：

$$(Bx)_1 = 2x_1 - x_2 = 2x_1 - x_2 - x_0 = (2 - 2\cos\theta)x_1$$
$$(Bx)_n = 2x_n - x_{n-1} = 2x_n - x_{n+1} - x_{n-1} = (2 - 2\cos\theta)x_n$$

把所有 n 个这样的等式合在一起，我们有

$$Bx = (2 - 2\cos\theta)x$$

■

9.1.5 稳定性分析：傅里叶方法

偏微分方程求解过程中的稳定性问题出现在几乎所有含有时间独立变量的问题中。因为我们对长时间区间上的解感兴趣，所以自然要讨论稳定性问题。上面分析显式方法中的稳定性使用的方法称为**矩阵方法**。另一个归于 von Neumann 的方法称为**傅里叶方法**。在此法中，我们试图求一个具有形式为

$$v_{jn} = e^{ij\beta h} e^{n\lambda k} \qquad (i = \sqrt{-1}) \tag{15}$$

的有限差分方程的解.（这里，我们用 j 代替 i 作为第 1 个下标使得它不会与复数 $i=\sqrt{-1}$混淆.）一旦这样做了以后，适当地选择 λ，考察当 $t\to\infty$或 $n\to\infty$时这个解的性态. 显然，这依赖于(15)式中的因子 $e^{\lambda nk}=(e^{\lambda k})^n$. 若 $|e^{\lambda k}|>1$，则解变成无界. 因为任何数值最终将被所有外来

621 的解所污染，因此这个无界的解将控制指数衰减的真解.

为什么我们考虑形如(15)式的解呢？有限差分方程的解应该具有与基本的偏微分方程的解相同的形式. 热传导方程(2)有形如 $u(x, t)=e^{-\pi^2 t}\sin(\pi x)$的解. 这就说明要考虑(15)式形式的解.

我们对(9)式给出的显式方法进行分析.（在此假定 $k>0$）. 把试验解(15)代入(9)式中，得到

$$e^{ij\beta h} e^{(n+1)\lambda k} = s e^{i(j-1)\beta h} e^{n\lambda k} + (1-2s) e^{ij\beta h} e^{n\lambda k} + s e^{i(j+1)\beta h} e^{n\lambda k}$$

消除因子 $e^{ij\beta h} e^{n\lambda k}$ 以后，结果是

$$\begin{aligned} e^{\lambda k} &= s e^{-i\beta h} + 1 - 2s + s e^{i\beta h} \\ &= 2s\cos\beta h + 1 - 2s \\ &= 1 - 2s(1 - \cos\beta h) \\ &= 1 - 4s\sin^2(\beta h/2) \end{aligned} \tag{16}$$

这里我们利用熟知的等式

$$e^{i\theta} = \cos\theta + i\sin\theta \qquad 1 - \cos\theta = 2\sin^2(\theta/2)$$

记得 $s=k/h^2$，其中 k 和 h 分别是 t 和 x 中的步长. 从(16)式显然可得 $e^{\lambda k}\leqslant 1$（只要 β 是实数）. 为了稳定性，我们还需要 $e^{\lambda k}\geqslant -1$，这导致限制条件

$$s\sin^2(\beta h/2) \leqslant 1/2$$

因为 $\sin^2(\beta h/2)$可能接近于 1，所以为了稳定性必有 $s\leqslant 1/2$.

习题 9.1

1. 证明：对固定的 n，显式算法中稳定性要求

$$s < \left(1 + \cos\frac{\pi}{n+1}\right)^{-1}$$

当 $n=10$ 时，k 的什么值是符合要求的？

2. 说明函数

$$u(x,t)=\sum_{n=1}^{N}c_n\exp(-n^2\pi^2 t)\sin n\pi x$$

是带有边界条件

$$\begin{cases}u(x,0)=\sum_{n=1}^{N}c_n\sin n\pi x\\ u(0,t)=u(1,t)=0\end{cases}$$

的热传导问题 $u_{xx}=u_t$ 的解.

622

3. 当 $h=10^{-2}$ 且 $0\leqslant t\leqslant 10$ 时，证明为了稳定性大约需要 10^7 个网格点. 当 $h=10^{-4}$ 时对应的数是什么?

计算机习题 9.1

利用显式方法的算法，求下列热传导问题的数值解

$$\begin{cases}u_{xx}=u_t\\ u(x,0)=\sin\pi x\\ u(0,t)=u(1,t)=0\end{cases}$$

在第1次实验中用 $h=0.1$，$k=0.005125$ 和 $M=200$. 比较计算解和精确解 $u(x,t)=\exp(-\pi^2 t)\sin(\pi x)$. 在第2次实验中，改变 k 为 0.006 以及 M 为 171.

9.2 抛物型方程：隐式方法

我们继续研究热传导模型问题. 记得

$$\begin{cases}u_{xx}=u_t & (t\geqslant 0,0\leqslant x\leqslant 1)\\ u(x,0)=g(x) & (0\leqslant x\leqslant 1)\\ u(0,t)=u(1,t)=0 & (t\geqslant 0)\end{cases}\tag{1}$$

正如 9.1 节中那样，导数用它的近似代替，且用 $v(x,t)$ 表示离散化问题的解. 现在考虑的有限差分方程是

$$\frac{1}{h^2}[v(x+h,t)-2v(x,t)+v(x-h,t)]=\frac{1}{k}[v(x,t)-v(x,t-k)]\tag{2}$$

利用 9.1 节中建立的记号，我们把(2)式记为

$$\frac{1}{h^2}[v_{i+1,j}-2v_{ij}+v_{i-1,j}]=\frac{1}{k}[v_{ij}-v_{i,j-1}]\tag{3}$$

这个式子与 9.1 节中的(7)式相比似乎仅仅表面上不同，但为了它的解需要一个不同类型的算法. 观察(3)式中的三项涉及 v 在 j 层(t)，并且只有一项涉及 v 在($j-1$)层($t-k$). 若 v 在($j-1$)层的网格点上已知，则仅通过解一个方程组就可从(3)式中算出在 j 层上的值. 所以我们利用 $s=k/h^2$ 把(3)式改写成如下形式

$$-sv_{i-1,j}+(1+2s)v_{ij}-sv_{i+1,j}=v_{i,j-1}\qquad(1\leqslant i\leqslant n)\tag{4}$$

如前，设 V_j 是向量

623

$$V_j=\begin{bmatrix}v_{1j}\\ v_{2j}\\ \vdots\\ v_{nj}\end{bmatrix}$$

为确定向量 V_j，把(4)式表示成一个 n 个方程的方程组，当然假定 V_{j-1} 已知. 这个方程组具有形式

$$AV_j = V_{j-1} \tag{5}$$

其中 A 是 $n\times n$ 矩阵

$$A = \begin{bmatrix} 1+2s & -s & 0 & \cdots & 0 \\ -s & 1+2s & -s & \cdots & 0 \\ 0 & -s & 1+2s & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1+2s \end{bmatrix} \tag{6}$$

形式上，方程(5)的解为

$$V_j = A^{-1}V_{j-1}$$

由此我们得到

$$V_j = A^{-1}(A^{-1}V_{j-2}) = A^{-1}(A^{-1}(A^{-1}V_{j-3})) = \cdots = A^{-j}V_0$$

因为 V_0 包含初值 $u(ih, 0)$，所以向量 V_0 已知. 根据 9.1 节中用过的相同理由，为算法的稳定性，我们要求 $\rho(A^{-1})<1$. 利用 9.1 节中(14)式的矩阵 B，我们看出

$$A = I + sB$$

由 9.1 节中的引理 1，A 的特征值是

$$\lambda_i = 1 + 2s(1-\cos\theta_i) \qquad \theta_i = \frac{i\pi}{n+1} \qquad (1\leqslant i \leqslant n) \tag{7}$$

因为这些显然满足 $\lambda_i>1$，所以 A^{-1} 的特征值位于区间(0，1)中，并且我们断定提出的方法对一切 h 和 k 的值是**稳定的**. 这个方法有时称为**全隐式方法**.

9.2.1 算法

现在系统地阐述执行全隐式方法的伪代码. 这个程序涉及其他两个程序，其中一个提供初值 $g(x)$. 另一个程序取名 tri；它的功能是解三对角线性方程组. 在 4.3 节末尾给出的伪代码是适合这里的. 在那个程序中，对角元素为 c_1，c_2，…，c_{n-1} 而次对角线元素为 a_1，a_2，…，a_{n-1}. 在右边的数是 b_1，b_2，…，b_n，而解记为 x_1，x_2，…，x_n. 因此，

624

tri(n,a,d,c,b;x)

输入 n，a，d，c，b 并提供输出 x. 在目前的应用中，a，d 和 c 是指定的常数值. 我们将利用初值 $v_i=g(ih)$ 代替 b. 程序 tri 提供向量 v 的下一个值并覆盖掉前面的值. 对角元 d_i 在 tri 中不变，所以在每个时间步必须重新初始化.

```
input n, k, M
h←1/(n+1)
s←k/h²
for i=1 to n do
    v_i←g(ih)
end do
t←0
```

```
output 0, t, (v1, v2, ..., vn)
for i=1 to n-1 do
    ci ← -s
    ai ← -s
end do
for j=1 to M do
    for i=1 to n do
        di ← 1+2s
    end do
    call tri(n, a, d, c, v; v)
    t ← jk
    output j, t, (v1, v2, ..., vn)
end do
```

9.2.2 克兰克-尼科尔森方法

可以把隐式和显式方法组合成一个包含参数 θ 的更一般的公式. 这个公式是

$$\frac{\theta}{h^2}(v_{i+1,j}-2v_{ij}+v_{i-1,j})+\frac{1-\theta}{h^2}(v_{i+1,j-1}-2v_{i,j-1}+v_{i-1,j-1})$$

$$=\frac{1}{k}(v_{ij}-v_{i,j-1}) \tag{8}$$

可以立即看出, 当 $\theta=0$ 时, 这个公式得到上节讨论的显式格式.（见 9.1 节中的(8)式.）当 $\theta=1$ 时, 公式化为上面讨论的隐式格式. 特殊情况 $\theta=\frac{1}{2}$ 导致用它的创造者 John Crank 和 Phyllis Nicolson 名字称呼的数值方法.

现在我们更详细地研究克兰克-尼科尔森方法. 有关的公式写成下列形式, 其中, 新的点(对应于 j)被放在左边而老的点(对应于 $j-1$)被放在右边.

$$-sv_{i-1,j}+(2+2s)v_{ij}-sv_{i+1,j}=sv_{i-1,j-1}+(2-2s)v_{i,j-1}+sv_{i+1,j-1} \tag{9}$$

625

这里 $s=k/h^2$. 如前, 引进包含元素 v_{ij}, $1\leqslant i\leqslant n$ 的向量 V_j, (9)式具有向量形式

$$(2I+sB)V_j=(2I-sB)V_{j-1} \tag{10}$$

其中 B 如前面一样, 见 9.1 节的(14)式. 根据熟知的理由, 如果

$$\rho[(2I+sB)^{-1}(2I-sB)]<1 \tag{11}$$

则保证获得方法的稳定性.

若 μ_1, μ_2, …, μ_n 是 B 的特征值, 则要求(11)变成

$$|(2+s\mu_i)^{-1}(2-s\mu_i)|<1 \tag{12}$$

因为 $\mu_i=2(1-\cos\theta_i)$, 所以可得 $0<\mu_i<4$. 接着作一点点代数运算就说明不等式(12)成立. 所以克兰克-尼科尔森方法对一切比值 $s=k/h^2$ 是稳定的.

当然, 稳定性不是这些方法中用来选择步长 h 和 k 的仅有的准则. 一般来说, 取的 h 和 k 越小, 离散化问题就更接近于模拟原来的微分方程. 我们需要的是一个定理, 它能保证在 $h\to 0$ 和 $k\to 0$ 时离散问题的解 $v(x, t)$ 收敛于原问题的解 $u(x, t)$. 这一结果是下面要讨论的内容.

9.2.3 分析

用等式

$$e_{ij}=u(x_i,t_j)-v(x_i,t_j) \tag{13}$$

定义网格点上的误差．我们把 $u(x_i,\ t_j)$ 和 $v(x_i,\ t_j)$ 缩写为 u_{ij} 和 v_{ij}．因此，$v_{ij}=u_{ij}-e_{ij}$．我们分析前一节中一维热传导方程的显式方法及其收敛性质．定义这个方法的等式是

$$v_{i,j+1}=s(v_{i-1,j}-2v_{ij}+v_{i+1,j})+v_{i,j} \tag{14}$$

这是 9.1 节的(9)式．照例，$s=k/h^2$．在(14)式中，我们用 $u-e$ 代替 v，得到

$$\begin{aligned}u_{i,j+1}-e_{i,j+1}&=s(u_{i-1,j}-2u_{ij}+u_{i+1,j})+u_{ij}\\&\quad-s(e_{i-1,j}-2e_{ij}+e_{i+1,j})-e_{ij}\end{aligned}$$

它可以重新整理成如下形式

626

$$\begin{aligned}e_{i,j+1}&=se_{i-1,j}+(1-2s)e_{ij}+se_{i+1,j}\\&\quad-s[u_{i-1,j}-2u_{ij}+u_{i+1,j}]+[u_{i,j+1}-u_{ij}]\end{aligned} \tag{15}$$

为简化此式中有括号的项，我们参照前面建立的数值微分公式，即，

$$f''(x)=\frac{1}{h^2}[f(x+h)-2f(x)+f(x-h)]-\frac{h^2}{12}f^{(4)}(\xi) \tag{16}$$

$$g'(t)=\frac{1}{k}[g(t+k)-g(t)]-\frac{k}{2}g''(\tau) \tag{17}$$

利用这些公式，从(15)式我们得到

$$\begin{aligned}e_{i,j+1}&=se_{i-1,j}+(1-2s)e_{i,j}+se_{i+1,j}\\&\quad-s\left[h^2u_{xx}(x_i,t_j)+\frac{h^4}{12}u_{xxxx}(\xi_i,t_j)\right]\\&\quad+\left[ku_t(x_i,t_j)+\frac{k^2}{2}u_{tt}(x_i,\tau_j)\right]\end{aligned} \tag{18}$$

现在利用 $sh^2=k$ 和 $u_{xx}=u_t$，(18)式可写成

$$\begin{aligned}e_{i,j+1}&=se_{i-1,j}+(1-2s)e_{ij}+se_{i+1,j}\\&\quad-kh^2\left[\frac{1}{12}u_{xxxx}(\xi_i,t_j)-\frac{s}{2}u_{tt}(x_i,\tau_j)\right]\end{aligned} \tag{19}$$

限制$(x,\ t)$在紧集 $S=\{(x,\ t):0\leqslant x\leqslant 1,\ 0\leqslant t\leqslant T\}$中．然后取

$$M=\frac{1}{12}\max|u_{xxxx}(x,t)|+\frac{s}{2}\max|u_{tt}(x,t)| \tag{20}$$

其中最大值是取遍一切$(x,\ t)\in S$．我们还定义误差向量

$$E_j=\begin{bmatrix}e_{1j}\\e_{2j}\\\vdots\\e_{nj}\end{bmatrix}$$

并取

$$\|E_j\|_\infty=\max_{1\leqslant i\leqslant n}|e_{ij}|$$

最后，我们假设 $1-2s\geqslant 0$. 于是由(19)式我们推断

$$\begin{aligned}|e_{i,j+1}| &\leqslant s|e_{i-1,j}|+(1-2s)|e_{ij}|+s|e_{i+1,j}|+kh^2M\\ &\leqslant s\|E_j\|_\infty+(1-2s)\|E_j\|_\infty+s\|E_j\|_\infty+kh^2M\\ &=\|E_j\|_\infty+kh^2M \qquad (21)\end{aligned}$$

627

因为(21)式右边不涉及 i，所以我们得到

$$\begin{aligned}\|E_{j+1}\|_\infty &\leqslant \|E_j\|_\infty+kh^2M\leqslant \|E_{j-1}\|_\infty+2kh^2M\leqslant\cdots\\ &\leqslant \|E_0\|_\infty+(j+1)kh^2M\end{aligned}$$

因为解 $v(x, t)$是从正确的初值开始的，所以有 $E_0=0$，并且

$$\|E_j\|_\infty\leqslant jkh^2M$$

现在设 $jk=t$，使得 $\|E_j\|_\infty$表示 t 层网格点上最大的误差．因为 $t\leqslant T$，所以

$$\|E_j\|_\infty\leqslant Th^2M=\mathcal{O}(h^2)$$

因此，倘若 $s=k/h^2\leqslant 1/2$ 且函数 u_{xxxx} 和 u_{tt} 是连续的，则当 $h\to 0$ 时，在任何固定的 t 层上最大的误差以 h^2 的速度收敛于 0.

9.2.4 小结

我们为得到一维热传导方程的数值解已讨论了三种方法：

方法	矩阵方程
显式	$V_{j+1}=(I-sB)V_j$
隐式	$(I+sB)V_j=V_{j-1}$
克兰克-尼科尔森	$(2I+sB)V_j=(2I-sB)V_{j-1}$

我们仅用两个涉及三对角方程组的子程序：

1. 计算 $y=Tx$.
2. 由 $Tx=b$ 解 x.

便可以容易地实施这些方法.

习题 9.2

1. 证明当 $r>0$ 时矩阵$(I+rB)^{-1}(I-rB)$的最大特征值是$(1-q)/(1+q)$，其中 $q=4r\sin^2[\pi/(2n+2)]$.
2. 证明：由(8)式定义的方法的稳定性条件是不等式 $s\leqslant(2-4\theta)^{-1}$，$0\leqslant\theta<\frac{1}{2}$，但是当$\frac{1}{2}\leqslant\theta\leqslant 1$ 时，对 s 没有限制.
3. 对全隐式方法进行收敛性分析．提示：在分析中的某一步，应该有下列形式的向量等式

$$(I+sB)E_j=E_{j-1}-C_j$$

其中 C_j 是一个向量，它含有下列形式的分量：

$$\frac{k^2}{2}u_{tt}(x_i,t_j+\theta_2k)+\frac{sh^4}{12}u_{xxxx}(x_i+\theta_1h,t_j)$$

628

4. 推广本节中的算法以便处理像下列问题中的终端条件：

$$\begin{cases}u_{xx}=u_t & (t\geqslant 0, 0\leqslant x\leqslant 1)\\ u(x,0)=g(x) & (0\leqslant x\leqslant 1)\\ u(0,t)=a(t) & (t\geqslant 0)\\ u(1,t)=b(t) & (t\geqslant 0)\end{cases}$$

计算机习题 9.2

1. 利用全隐式方法在单位正方形上求解下列热传导问题

$$\begin{cases} u_{xx} = u_t \\ u(x,0) = (x - x^2)e^x \\ u(0,t) = u(1,t) = 0 \end{cases}$$

建议值是 $n=20$，$M=50$，$k=0.05$.

2. 对下列问题

$$\begin{cases} u_{xx} - \alpha u_t = f(x) & (0 < x < L, 0 < t) \\ u(x,0) = g(x) & (0 < x < L) \\ u(0,t) = u(L,t) = 0 & (0 < t) \end{cases}$$

编写且测试一个执行克兰克-尼科尔森过程的程序.

3. 利用两个涉及三对角方程组运算的程序比较显式、隐式和克兰克-尼科尔森方法，并利用计算机习题 9.1.1 对它们进行测试.

9.3 定常问题：有限差分法

两个变量的**拉普拉斯方程**

$$u_{xx} + u_{yy} = 0 \tag{1}$$

是一个典型的时间不是变量的偏微分方程. 在这个方程中，u 是(x, y)的函数，与(1)式有关的一个具体的物理问题也包含指定 xy 平面内的一个求解的区域 Ω，并对 u 设置边界条件(例如假定在 Ω 的边界上 u 的值或法向导数). 至于$\mathbb{R}^2$ 中的区域 Ω，假定 Ω 是一个开集. 它的边界用 $\partial\Omega$ 表示，而它的闭包用 $\overline{\Omega}$ 表示. 因此，$\overline{\Omega}=\Omega\cup\partial\Omega$.

9.3.1 狄利克雷问题

在热学、电学和许多其他物理学分支的研究中发生的问题是**狄利克雷问题**. 在二维形式中，在$\mathbb{R}^2$ 中指定一个开域 Ω. 定义在 Ω 边界上的函数 g 也是已知的. 然后，我们求在闭包 $\overline{\Omega}$ 上的连续函数 u，在 Ω 中满足拉普拉斯方程且在边界上等于 g. 这些可归纳为

629

$$\begin{cases} u_{xx} + u_{yy} = 0 & \text{在 } \Omega \text{ 内} \\ u(x,y) = g(x,y) & \text{在 } \partial\Omega \text{ 上} \\ u \text{ 在 } \overline{\Omega} \text{ 上连续} \end{cases} \tag{2}$$

若 Ω 服从某个适度的限制并且 g 是连续的，则可以证明狄利克雷问题有唯一解.

为了说明某些数值方法，我们将在一个正方形区域上考虑狄利克雷问题. 这个问题可用分离变量和傅里叶级数的分析方法求解. 然而，对其他的区域，通常需要本节和下节中讨论的数值方法. 此外，还应该强调，即使能给出一个无穷级数形式的解，一个涉及问题离散化的数值解可能更可取.

在要说明的问题中，区域是开的单位正方形

$$\Omega = \{(x,y) : 0 < x < 1, 0 < y < 1\}$$

边界函数 g 暂时是任意的. 在计算机程序中它的值由一个适当的子程序来提供.

9.3.2 有限差分

数值求解问题(2)的一个方法是利用有限差分逼近导数. 可用 9.2 节中熟知的公式，即

$$f''(x) = \frac{1}{h^2}[f(x+h) - 2f(x) + f(x-h)] + \mathcal{O}(h^2) \tag{3}$$

首先，在正方形的 $\overline{\Omega}$ 中建立网格点的网络：

$$(x_i, y_j) = (ih, jh) \qquad (0 \leqslant i, j \leqslant n+1) \qquad h = \frac{1}{n+1} \tag{4}$$

注意两个变量使用相同的步长．其次，在网格点(x_i, y_j)上的微分方程(1)用它在这些点上有限差分模拟代替，它是

$$\frac{1}{h^2}[v_{i-1,j} - 2v_{ij} + v_{i+1,j}] + \frac{1}{h^2}[v_{i,j-1} - 2v_{ij} + v_{i,j+1}] = 0$$

或

$$4v_{ij} - v_{i-1,j} - v_{i+1,j} - v_{i,j-1} - v_{i,j+1} = 0 \tag{5}$$

这里，打算使 v_{ij} 逼近 $u(x_i, y_j)$．因为我们要区别下面两个不同的问题的解：第一个是**离散问题**(即有限差分方程)；第二个是**连续问题**(即原偏微分方程)，所以使用不同的变量．

630

当 $i=0$ 或 $n+1$ 及 $j=0$ 或 $n+1$ 时，v_{ij} 的值已知，因为这些是问题中规定的边界值(由函数 g 给出)．因而(5)式中的某些项 v_{ij} 可能已知而其他一些项未知．因为所有已知值已移到右边，所以，实际上这意味着我们将求解一个非齐次的线性方程组．取一个简单情况，设 $n=3$．对每个内部网格点有一个类型(5)的等式．在图 9-4 中，用大的点描绘这些网格点．

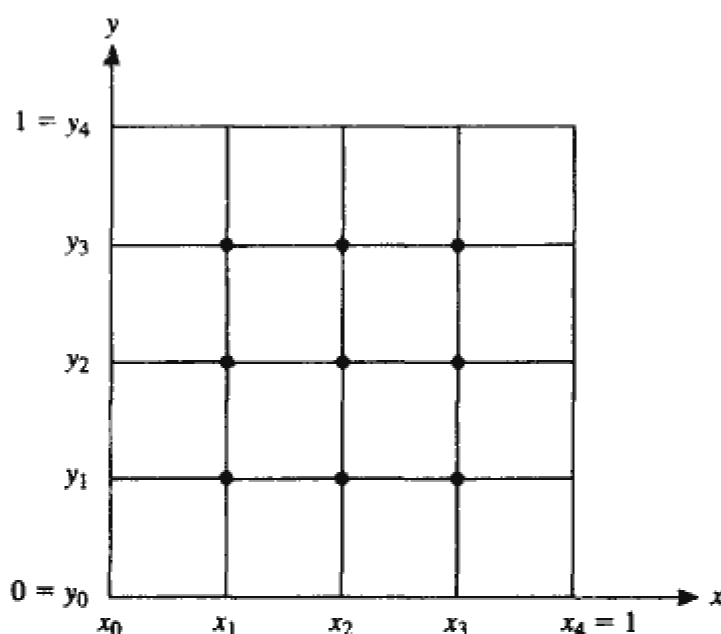


图 9-4 单位正方形网格

这个问题中的未知量可用许多方式排序(对应于网格点的不同次序)．我们选择人们通称的**自然次序**：

$$v = [v_{11}, v_{21}, v_{31}, v_{12}, v_{22}, v_{32}, v_{13}, v_{23}, v_{33}]$$

同样，9 个线性方程可用许多方式排序．我们决定按(5)式和它的中心点(x_i, y_j)对应起来排序，然后像对点排序那样对方程排序．结果如下，

$$\begin{aligned}
4v_{11} - v_{21} - v_{12} \qquad\qquad &= v_{10} + v_{01} \\
4v_{21} - v_{11} - v_{31} - v_{22} \qquad &= v_{20} \\
4v_{31} - v_{21} - v_{32} \qquad\qquad &= v_{30} + v_{41} \\
4v_{12} - v_{11} - v_{22} - v_{13} \qquad &= v_{02} \\
4v_{22} - v_{21} - v_{12} - v_{32} - v_{23} &= 0
\end{aligned}$$

$$
\begin{aligned}
4v_{32}-v_{31}-v_{22}-v_{33} &= v_{42} \\
4v_{13}-v_{12}-v_{23} &= v_{03}+v_{14} \\
4v_{23}-v_{22}-v_{13}-v_{33} &= v_{24} \\
4v_{33}-v_{32}-v_{23} &= v_{43}+v_{34}
\end{aligned}
$$

其中所有的已知量已移到等式的右边．这个方程组具有形式 $Av=b$，其中矩阵 A 是 9×9 的，81 个元素中只有 33 个是非零的．系数矩阵为

631

$$
A=\begin{bmatrix}
\begin{bmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & -4 \end{bmatrix} & \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} & \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} & \begin{bmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix} & \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \\
\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} & \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} & \begin{bmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix}
\end{bmatrix}
$$

在这个矩阵中，已指出基本的块结构．我们可用显而易见的符号写出

$$
A=\begin{bmatrix} T & -I & 0 \\ -I & T & -I \\ 0 & -I & T \end{bmatrix}
$$

用前面的特殊情况作为一个模型，我们进入到一般情况，这里 n 可以是任意的整数．内部网格点数为 n^2，每一个点有一个对应的未知值 $u_{ij}=u(x_i,\ y_j)$．函数 u 是方程(2)的解．离散化以后，得到一个确定 n^2 个未知量 v_{ij} 的线性方程组，其中 $1\leqslant i\leqslant n$ 且 $1\leqslant j\leqslant n$．

9.3.3 算法

因为每个方程至多包含 5 个未知数，所以方程组是稀疏的(总共 n^2 个元素集合中至多有 $5n$ 个非零元)．像 4.6 节中所讨论的那些迭代法在这种情况可能是十分有效的．已知的边界值是

$$
v_{ij}=g(x_i,y_j),\text{当 } i \text{ 或 } j \text{ 等于 } 0 \text{ 或 } n+1 \text{ 时} \tag{6}
$$

因此，在算法中，元素为 v_{ij} 的整个数组允许值 $1\leqslant i,\ j\leqslant n+1$，并包含 $(n+2)^2$ 个元素．作为说明，我们提出高斯-赛德尔迭代法．对这个过程，对应于点 $(x_i,\ y_j)$(5)式可写成形式

$$
v_{ij}=(v_{i-1,j}+v_{i+1,j}+v_{i,j-1}+v_{i,j+1})/4 \tag{7}
$$

此式用来更新 v_{ij}．当使用此式时，从右边得到的值替换 v_{ij} 的旧值．如果我们仅当 $1\leqslant i,\ j\leqslant n$ 时仔细地使用(7)式，则只有内部的网格点涉及这一更新．边界上的网格点只出现在(7)式的右边．计算的要点如下：

1. 把边界值放到适当的 v_{ij} 上．
2. 对内部的结点提供合适的初值．
3. 执行 M 步高斯-赛德尔迭代．

632

第 1 项工作的一段伪代码是：

```
for i=0 to n+1 do
    v_i0 = g(x_i, 0); v_i,n+1 = g(x_i, 1)
    v_n+1,i = g(1, y_i); v_0i = g(0, y_i)
end do
```

第 3 项工作的伪代码是：

```
for k=1 to M do
    for j=1 to n do
        for i=1 to n do
            v_ij = (v_i-1,j + v_i+1,j + v_i,j-1 + v_i,j+1)/4.0
        end do
    end do
end do
```

注意这个算法中的嵌套循环按选择的次序(自然次序)更新向量 v.

对一个具体情况，我们用公式

$$g(x,y) = 10^{-4}\sin(3\pi x)\sin(3\pi y)$$

作为已知的函数 g. 在计算机程序中，这个函数 g 仅对正方形边界点使用. 但是 g 是调和的，故 g 也是问题的解. 因而能够计算近似解 v_{ij} 并与真解 $u(x_i,\ y_j)=g(x_i,\ y_j)$ 比较. 设 $n=18$，使得 A 是 324×324. 因为高斯-赛德尔方法不需要这个矩阵，所以这个矩阵不存储在计算机中. 代之，我们使用(7)式，并且只需要有 324 个分量 v_{ij} $(0\leqslant i\leqslant 18$ 且 $0\leqslant j\leqslant 18)$ 的向量 v. 当用双精度编程时，经 200 次迭代后，得到的误差满足不等式

$$|v_{ij} - u(x_i,y_j)| < 0.345\times 10^{-7}$$

习题 9.3

1. 设∇^2表示拉普拉斯算子：$\nabla^2 u = u_{xx}+u_{yy}$. 证明：问题

$$\begin{cases}\nabla^2 u = f & \text{在 } \Omega \text{ 内}\\ u = 0 & \text{在 } \partial\Omega \text{ 上}\end{cases}$$

可用下面三步求解：

ⅰ. 求 g 使$\nabla^2 g = f$.

ⅱ. 利用 $-g$ 作为边界值在 Ω 中求解狄利克雷问题.

ⅲ. 把 g 加到在第 ⅱ 步中得到的函数上求出 u.

2. (续)描述等价于

$$\begin{cases}\nabla^2 u = (6-3x^2)\sin y - y\cos x & \text{在 } \Omega \text{ 内}\\ u = 0 & \text{在 } \partial\Omega \text{ 上}\end{cases}$$

的狄利克雷问题.

3. 平面上对点排序的一个方法是**字典次序**. 这是参照字典中使用的次序. 在字典中，我们首先放置用“a”开始的所有单词，在这些单词中，我们按第 2 个字母排序单词. 相同的情况，我们看第 3 个字母，等等. 因此，在平面上，字典次序定义为 633

$$(x,y)\leqslant(u,v),\text{若}(x<u)\text{或}(x=u\text{ 且 }y\leqslant v)$$

若网格点及其方程按字典次序排列时，在模型问题中产生怎样的矩阵？

4. 对 n^2 个内部网格点的模型问题(2)，一般说来有多少方程会是齐次的？有多少方程会有 5 个非零系数在左

边？有多少方程会有 4 个非零系数在左边？有多少方程会有 3 个非零系数？矩阵中非零系数的精确个数是多少？

5. 当边界值除了 $u(x,\ 0)=\sin 4\pi x$ 外均为 0 时，利用分离变量法和傅里叶级数求单位正方形上狄利克雷问题的解析解.

6. 设 ∇^2 是二维拉普拉斯算子，而 δ 是由下式定义的离散的拉普拉斯算子

$$(\delta f)(x,y) = [f(x+1,y) - 2f(x,y) + f(x-1,y)]$$
$$= [f(x,y+1) - 2f(x,y) + f(x,y-1)]$$

证明或否定：对每个 $(x,\ y)$，存在一个 $(\xi,\ \eta)$ 使得

$$(\delta f) = (\nabla^2 f)(\xi,\eta)$$

它的一维形式是什么？

计算机习题 9.3

1. 对课文中提出的求解正方形区域上狄利克雷问题的过程编写程序. 点数、迭代数和打印次数应该用容易改变的参数控制. 边界值应该用一个子程序提供. 利用

$$g(x,y) = 4xy(x-y)(x+y)$$

测试你的程序. 比较连续问题和离散问题的解.

2. 对课本中的例子编写程序并看看你的计算机产生的结果是否类似于那些引用的结果.
3. (续)利用 SOR(逐次超松弛)法代替高斯-赛德尔方法修正程序.

9.4 定常问题：伽辽金法

伽辽金法广泛地用于那些需要确定未知函数的问题. 当然，微分方程和积分方程属于这个范畴. 当方法应用于任意的线性问题时，我们先概要叙述原理. 然后，用求解长方形区域上的狄利克雷问题的一个数值例子加以说明.

634

9.4.1 伽辽金法

假定我们遇到下列形式的问题

$$Lu = f \tag{1}$$

其中 L 是一个线性算子，f 是一个已知的函数，而 u 是由方程确定的一个函数. 在**伽辽金法**中，我们选取一组**基函数**或**试验函数** $u_1,\ u_2,\ \cdots,\ u_n$. 然后，打算用这些基函数的一个适当的线性组合去解方程(1). 取

$$u = \sum_{j=1}^{n} c_j u_j$$

并利用 L 的线性性，得到

$$\sum_{j=1}^{n} c_j L u_j = f \tag{2}$$

在典型的情况下，因为 f 一般不位于由函数 Lu_j 张成的向量空间中，所以这个方程**不相容**. 因此，我们近似地求解方程(2)，从而得到方程(1)的近似解. 方程(2)的近似求解可按许多不同的准则执行，每个准则导致不同的近似解 u. 最自然的方法是选择 $c_1,\ c_2,\ \cdots,\ c_n$ 使某种范数达到极小：

$$\left\| \sum_{j=1}^{n} c_j L u_j - f \right\| = \min \tag{3}$$

这是最佳逼近中的一个问题．我们用函数 Lu_j 生成的子空间中最近的元素来逼近 f．因为可以利用正交投影方法，所以这在内积空间中是相对容易的．

得到方程(2)近似解的一个非常一般的方法是选择一组线性函数 $\phi_1, \phi_2, \cdots, \phi_n$ 并且强加条件

$$\phi_i\left(\sum_{j=1}^{n} c_j L u_j - f\right) = 0 \qquad (1 \leqslant i \leqslant n) \tag{4}$$

利用函数的线性性，这个条件变成

$$\sum_{j=1}^{n} \phi_i(Lu_j) c_j = \phi_i(f) \qquad (1 \leqslant i \leqslant n) \tag{5}$$

(5)式是 n 个未知数 c_j 的 n 个线性方程的方程组．若泛函是由下式定义的点赋值泛函

$$\phi_i(v) = v(x_i) \qquad (1 \leqslant i \leqslant n) \tag{6}$$

635

则上面列举的方法称为**配置法**．我们把前述方法的所有形式称为**伽辽金法**．经典的伽辽金法是(5)式在希尔伯特空间中的一个特殊情况，其中 $\phi_i(v) = \langle u_i, v \rangle$．因此，此时求解的方程是

$$\sum_{j=1}^{n} c_j \langle u_i, Lu_j \rangle = \langle u_i, f \rangle \qquad (1 \leqslant i \leqslant n)$$

在 8.10 节中，配置法用于求解涉及微分方程的两点边值问题．这里我们将说明狄利克雷问题的伽辽金法．所选择的例子也是一个能用分离变量和傅里叶级数求解的．

9.4.2 狄利克雷问题

考虑下述平面内的一个开放矩形

$$\Omega = \{(x, y) : |x| < 1, |y| < 2\} \tag{7}$$

我们寻找一个定义在 $\overline{\Omega}$ 上的连续函数 u，使得

$$\begin{cases} \nabla^2 u = 0 & \text{在 } \Omega \text{ 内} \\ u(x, y) = x^2 + y^2 & \text{在 } \partial\Omega \text{ 上} \end{cases} \tag{8}$$

这个问题可认为是 $Lu = f$ 的一种形式；用等式简单地定义 L 和 f

$$Lu = \begin{bmatrix} \nabla^2 u \\ u \mid \partial\Omega \end{bmatrix} \qquad f(x, y) = \begin{bmatrix} 0 \\ x^2 + y^2 \end{bmatrix} \tag{9}$$

记号 $u \mid S$ 表示函数 u 在集合 S 上的限制．它是一个线性运算，因为

$$(\alpha u + \beta v) \mid S = \alpha(u \mid S) + \beta(v \mid S)$$

为了使用伽辽金法，选择一组适当的基函数是必要的．对下面讨论的问题，我们选择满足问题的齐次部分的函数 $u_1, u_2, \cdots, u_n$．即我们选择 u_i 使得在 Ω 内 $\nabla^2 u_i = 0$．这种函数称为在 Ω 内是**调和**的．利用任何复变量解析函数的实部和虚部是调和的原理的优点，可以利用大量的调和函数．前几个幂函数 z^k 提供一组方便的调和多项式：

$$z^0 = 1$$

$$z^1 = x + \mathrm{i}y$$

$$z^2 = (x^2 - y^2) + (2xy)\mathrm{i}$$

636

$$z^3 = (x^3 - 3xy^2) + (3x^2y - y^3)\mathrm{i}$$
$$z^4 = (x^4 - 6x^2y^2 + y^4) + (4x^3y - 4xy^3)\mathrm{i}$$
$$z^5 = (x^5 - 10x^3y^2 + 5xy^4) + (5x^4y - 10x^2y^3 + y^5)\mathrm{i}$$
$$z^6 = (x^6 - 15x^4y^2 + 15x^2y^4 - y^6) + (6x^5y - 20x^3y^3 + 6xy^5)\mathrm{i}$$

因为我们的问题关于坐标轴和原点具有对称性，所以选择有相同对称性的 z^k 的实部和虚部.这就限定了 k 为偶数的 z^k 的实部. 因此，若设 $n=4$，则可使用基函数

$$u_1(x,y) = 1$$
$$u_2(x,y) = x^2 - y^2$$
$$u_3(x,y) = x^4 - 6x^2y^2 + y^4$$
$$u_4(x,y) = x^6 - 15x^4y^2 + 15x^2y^4 - y^6$$

对基函数的这种选择,函数 $u = \sum_{j=1}^{4} c_j u_j$ 显然将自动满足 $\nabla^2 u = 0$,并且应该选择系数 c_j 使边界条件近似地满足.应当近似求解的方程是

$$\sum_{j=1}^{4} c_j u_j(x,y) = x^2 + y^2 \quad 在\ \partial\Omega 上 \tag{10}$$

由于对称性，所以只考虑位于第一象限中的 $\partial\Omega$ 部分就足够了.

为说明配置法，我们选择 $\partial\Omega$ 上的 4 个点，即(0, 2)，(1, 0)，(1, 1)和(1, 2). 要求(10)式在这 4 个点上成立. 然后导出下列线性方程组：

$$\begin{bmatrix} 1 & -4 & 16 & -64 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & -4 & 0 \\ 1 & -3 & -7 & 117 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} 4 \\ 1 \\ 2 \\ 5 \end{bmatrix}$$

这个方程组的近似解是

$$c = (1.826\,1, -0.787\,0, -0.043\,48, 0.004\,348)^{\mathrm{T}}$$

因此 $\sum_{j=1}^{4} c_j u_j$ 和函数 $x^2 + y^2$ 之间在矩形边界上最大的偏差近似等于 0.074.

在前面的例子中，我们保留基函数组但以另外的方法“解”方程(10). 在 Ω 的边界上选择一组 m 个点(x_i, y_i)并尝试关于 c_j 极小化表达式

637

$$\sum_{i=1}^{m} \Big[\sum_{j=1}^{4} c_j u_j(x_i, y_i) - (x_i^2 + y_i^2) \Big]^2 \tag{11}$$

因为可以使用 $m=4$，所以这个过程包括配置作为一个特殊情况. 但是通过取更多的点，我们可以做极小化平方 ℓ^2 范数

$$\int_{\partial\Omega} \Big[\sum_{j=1}^{4} c_j u_j(x,y) - (x^2 + y^2) \Big]^2 \mathrm{d}x\mathrm{d}y \tag{12}$$

作为这个方法的实际试验，我们选择 $\partial\Omega$ 的第一象限中的 76 个等距点. 极小化表达式(11)变成具有 76 个方程和 4 个未知量的最小二乘矩阵问题. 计算的结果是

$$c = (1.821\,6, -0.781\,1, -0.044\,58, 0.004\,052)^{\mathrm{T}}$$

76 个点中的最大偏差是 0.049. 使用的 76 个点是(1, $i/25$), $0\leqslant i\leqslant 50$ 和($i/25$, 2), $0\leqslant i\leqslant 24$.

最后，可以尝试极小化

$$\max_{1\leqslant i\leqslant m}\left|\sum_{j=1}^{4}c_ju_j(x_i,y_i)-(x_i^2+y_i^2)\right| \tag{13}$$

这称为 4 个未知量的 m 个线性方程的超定方程组的**极小化极大解**. 对大的 m，这个过程做极小化一致范数

$$\max_{(x,y)\in\partial\Omega}\left|\sum_{j=1}^{4}c_ju_j(x,y)-(x^2+y^2)\right| \tag{14}$$

这个问题或它的离散模拟(13)，可以用列梅兹算法(见 6.9 节)求解. 当表达式(13)用和最小二乘解中同样的一组 76 个点极小化时，c 的值变成

$$c=(1.807\,2,-0.795\,0,-0.040\,0,0.003\,692)^{\mathrm{T}}$$

而 76 个点中的最大偏差是 0.033. 用这个更加复杂的方法得到解所花费的额外努力几乎不能在精确度方面有最低限度的改善. 增加基函数数目应该是比较好的方法.

9.4.3 泊松方程

涉及**泊松方程**的边值问题是经常遇到的. 典型的情况可能像下面那样：

$$\begin{cases}\nabla^2 w=f & \text{在 }\Omega\text{ 内}\\ w=g & \text{在 }\partial\Omega\text{ 上}\end{cases} \tag{15}$$

638

着手解决此问题的一个方法是把它分成两个较简单的问题：

$$\begin{cases}\nabla^2 v=0 & \text{在 }\Omega\text{ 内}\\ v=g & \text{在 }\partial\Omega\text{ 上}\end{cases}\qquad\begin{cases}\nabla^2 u=f & \text{在 }\Omega\text{ 内}\\ u=0 & \text{在 }\partial\Omega\text{ 上}\end{cases}$$

在得到 u 和 v 后,原问题的解就是 $w=u+v$. 这两个较简单的问题的优点是,每个问题都有一个齐次的部分. 在伽辽金法中利用了这个特征. 因此,在求解涉及 u 的问题中,可选择在 Ω 的边界上是 0 的基函数. 若 $u_1,u_2,\cdots,u_n$ 具有这个性质,则对它们任意的线性组合这个性质同样成立. 其次,我们取 $u=\sum_{j=1}^{n}c_ju_j$ 并试图通过选择系数 c_j 使等式 $\nabla^2u=f$ 成立. 这导致等式

$$\sum_{j=1}^{n}c_j\nabla^2u_j=f \tag{16}$$

通常在这些方法中，任何选择的系数都不能使这个等式成立，并且只能找到近似解.

9.4.4 瑞利-里茨方法

求解问题

$$\begin{cases}\nabla^2 u=f & \text{在 }\Omega\text{ 内}\\ u=0 & \text{在 }\partial\Omega\text{ 上}\end{cases} \tag{17}$$

的另一种方法是**瑞利-里茨方法**. 区域 Ω 保持固定，我们在一个内积空间 V 中工作，V 中的元素是函数 u 使得在 Ω 中 u_{xx} 和 u_{yy} 是连续的并且使得在 Ω 的边界上 $u(x,\ y)=0$. 假定 f 在 Ω 中是连续的. 所以，问题的解应该在空间 V 中. V 中的内积用下式定义

$$\langle u,v\rangle=\iint_{\Omega}u(x,y)v(x,y)\,\mathrm{d}x\mathrm{d}y \tag{18}$$

定理1(拉普拉斯算子定理) 算子$-\nabla^2$在内积空间V上是自伴(或对称)和正定的.

证明 用等式

$$\langle -\nabla^2 u, v\rangle = \langle u, -\nabla^2 v\rangle \tag{19}$$

表达具有自伴(或对称)性质. 为证明(19)式，我们需要平面上的格林定理，该定理叙述了对适当的函数和适当的区域有

[639]

$$\iint_\Omega (P_x + Q_y)\mathrm{d}x\mathrm{d}y = \int_{\partial\Omega}(P\mathrm{d}y - Q\mathrm{d}x)$$

使用格林定理，且利用u和v在$\partial\Omega$上为0这一事实，我们可以如下计算：

$$\begin{aligned}\langle \nabla^2 u, v\rangle &= \iint_\Omega (u_{xx} + u_{yy})v\mathrm{d}x\mathrm{d}y \\ &= \iint_\Omega [(u_x v)_x + (u_y v)_y - u_x v_x - u_y v_y]\mathrm{d}x\mathrm{d}y \\ &= \int_{\partial\Omega}(u_x v\mathrm{d}y - u_y v\mathrm{d}x) - \iint_\Omega (u_x v_x + u_y v_y)\mathrm{d}x\mathrm{d}y \\ &= -\iint_\Omega (u_x v_x + u_y v_y)\mathrm{d}x\mathrm{d}y\end{aligned}$$

在这个计算中最后的表达式包含的u和v是对称的，所以

$$\langle \nabla^2 u, v\rangle = \langle u, \nabla^2 v\rangle \tag{20}$$

关于$-\nabla^2$的正定性，我们必须证明若$u\in V$且$u\neq 0$，则

$$\langle -\nabla^2 u, u\rangle > 0$$

前面的计算指出

$$\langle -\nabla^2 u, u\rangle = \iint_\Omega [(u_x)^2 + (u_y)^2]\mathrm{d}x\mathrm{d}y \tag{21}$$

这个积分的值一定是非负的，仅有的问题是对不为0的函数u它能否为0. 若积分值为0，则在Ω中$u_x = u_y = 0$. 因此，u同时是只有y的函数和只有x的函数. 于是，u在Ω中必然是常数. 但作为V的一个元素，u在$\partial\Omega$上必须为0. 因此，$u=0$. ■

因为算子$-\nabla^2$在V上是对称正定的，所以一个新的内积可定义为

$$[u, v] \equiv \langle -\nabla^2 u, v\rangle = \iint_\Omega (u_x v_x + u_y v_y)\mathrm{d}x\mathrm{d}y \tag{22}$$

这个新的内积通过下列等式导出空间V中的一个新范数

$$\| u \| = [u, u]^{1/2}$$

在瑞利-里茨方法中,我们在V中选取基函数$u_1, u_2, \cdots, u_n$并试图确定$c_1, c_2, \cdots, c_n$使$\sum_{j=1}^{n} c_j u_j$按范数$\|\cdot\|$尽可能接近于真解.因为内积空间的工具是可利用的,所以我们知道确定系数c_j的正规方程.因此,如果u是真解,则正交性条件必然成立:

[640]

$$u - \sum_{j=1}^{n} c_j u_j \perp u_i \qquad (1 \leqslant i \leqslant n)$$

这立即导致方程

$$\sum_{j=1}^{n} c_j [u_j, u_i] = [u, u_i] \qquad (1 \leqslant i \leqslant n)$$

这些方程可利用定义(22)及关系

$$[u,u_i]=\langle -\nabla^2 u,u_i\rangle=\langle -f,u_i\rangle \qquad (1\leqslant i\leqslant n)$$

简化，结果得到正规方程组

$$\sum_{j=1}^{n}c_j[u_j,u_i]=-\langle f,u_i\rangle \qquad (1\leqslant i\leqslant n) \tag{23}$$

我们看到在方程组(23)中没有出现未知函数 u.

9.4.5 有限元素法

当伽辽金法使用的基函数是分段多项式时，这个方法称为**有限元素法**. 在这个方法的一种表现形式中，区域 Ω 假定是多边形并符合图 9-5 所示那样的**三角剖分**.

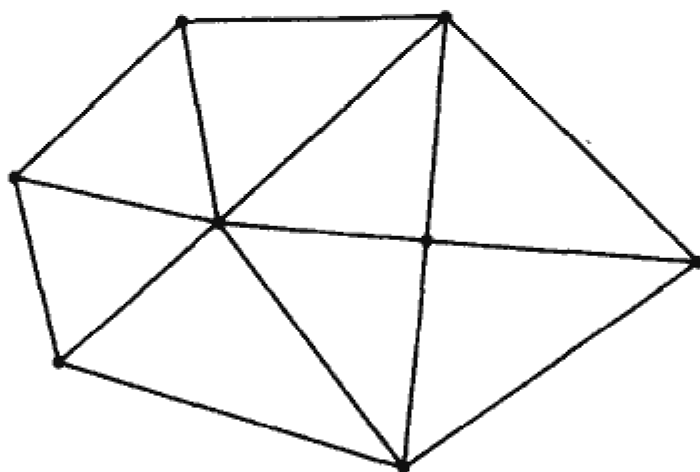


图 9-5 适当的三角剖分

(x, y)的线性函数具有形式 $ax+by+c$. 通过指定三角形的三个顶点上的函数值可以在每个三角形上定义一个这样的函数. 因为在任意适当的三角剖分中三角形共边，所以用这样的方法得到的分段线性函数是连续的. 因此，我们不允许像图 9-6 所示的那种三角剖分，因为通过指定每个三角形顶点上的任意值得到的分段线性函数可能不是连续的. 适当的三角剖分的法则是某个三角形的任何顶点必须是这个点所属的每个三角形的顶点. 在图 9-6 中点 e 是$\triangle abe$ 的一个顶点并且属于$\triangle dbc$，但它不是$\triangle dbc$ 的顶点.

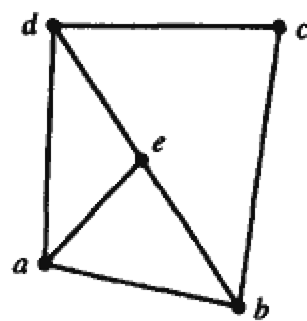


图 9-6 不适当的三角剖分

有限元素法已讨论了如此众多的问题，我们建议读者查询关于这些问题的专门文献. 例如，参见 Becker，Carey and Oden[1981]、Mitchell and Wait[1977]、Oden[1972]、Oden and Reddy[1976]、Strang and Fix[1973]、Vichnevetsky[1981]、Wait and Mitchell[1986]以及 Zienkiewicz and Morgan[1983].

习题 9.4

1. 证明：若 w 是 z 的一个解析函数($w=u+\mathrm{i}v$, $z=x+\mathrm{i}y$)，则 u 和 v 是调和的. 提示：利用柯西–黎曼方程.
2. 设 $z^n=u_n+\mathrm{i}v_n$. 证明 u_n 和 v_n 可由下列公式递归生成

$$\begin{cases} u_0=1 & v_0=0 \\ u_{n+1}=xu_n-yv_n & v_{n+1}=xv_n+yu_n \end{cases}$$

3.（续）证明：对每个偶数 n，u_n 和 v_n 有性质

$$u_n(-x,\ y)=u_n(x,\ -y)=u_n(x,\ y) \qquad v_n(-x,\ y)=v_n(x,\ -y)=-v_n(x,\ y)$$

4. 证明：如果由(7)式和(8)式定义的狄利克雷问题有解，则它有一个满足上面第3题的对称条件的解.

5. 波动方程的狄利克雷问题通常是不可解的. 例如，说明没有函数在单位正方形内满足 $u_{xy}=0$ 并取边界值 $u(x,\ 0)=x$，$u(0,\ y)=y$，$u(1,\ y)=1$.

6. 证明：若 $u_0+\mathrm{i}v_0$ 是解析的，则 $u_n+\mathrm{i}v_n$ 同样是解析的，其中这些函数由下列公式递归生成

$$u_{n+1}=xu_n-yv_n \qquad v_{n+1}=xv_n+yu_n$$

7. 推导基函数 $u_4(x,\ y)$ 和 $u_5(x,\ y)$ 的方程.

计算机习题 9.4

对课本中的例子编写程序并验证它们的正确性.（当然，由于用不同字长的计算机，结果将产生差别.）

9.5 一阶偏微分方程：特征线法

641～642

正如在常微分方程的理论中那样，高阶偏微分方程可以用一阶偏微分方程组替代. 下面用例子来说明如何进行.

9.5.1 一阶方程组

例 1 把9.1节中考虑的一维热传导方程

$$u_{xx}=u_t \tag{1}$$

化为一阶方程的方程组.

解 引进变量 $v=u_x$，我们得到等价的方程组：

$$\begin{cases} v_x-u_t=0 \\ u_x-v=0 \end{cases} \tag{2}$$

■

例 2 说明三维热传导方程

$$u_{xx}+u_{yy}+u_{zz}=u_t \tag{3}$$

可按照例1中相同的方式处理.

解 引进变量 $u^{(1)}=u$，$u^{(2)}=u_x$，$u^{(3)}=u_y$，$u^{(4)}=u_z$，我们得到一个等价的一阶方程组

$$\begin{cases} u_x^{(2)}+u_y^{(3)}+u_z^{(4)}-u_t^{(1)}=0 \\ u_x^{(1)}=u^{(2)} \\ u_y^{(1)}=u^{(3)} \\ u_z^{(1)}=u^{(4)} \end{cases} \tag{4}$$

■

例 3 对具有两个独立变量的一般二阶偏微分方程

$$F(x,y,u,u_x,u_y,u_{xx},u_{xy},u_{yy})=0 \tag{5}$$

重复例1的工作.

解 这个方程等价于下列一阶方程组：

$$\begin{cases} F(x,y,u,v,w,v_x,v_y,w_y)=0 \\ u_x=v \\ u_y=w \end{cases} \tag{6}$$

■

9.5.2 特征曲线

我们现在专注于偏微分方程的**特征曲线**，或简称**特征线**的概念．对给定方程的一类特征曲线是一条曲线，在这条曲线上解是常数．为说明这个可能性，考虑下列一阶方程(它与二阶波动方程密切相关)：

$$u_x + cu_y = 0 \tag{7}$$

[643]

设给定 xy 平面上的一条曲线为函数 $y=y(x)$ 的图像．沿这条曲线，u 的值仅仅是 x 的函数 $u(x, y(x))$．假设曲线是一条表达式为常数的曲线．则下列条件必须满足：

$$0 = \frac{\mathrm{d}}{\mathrm{d}x}u(x,y(x)) = u_x + u_y\frac{\mathrm{d}y}{\mathrm{d}x} \tag{8}$$

于是，我们寻找的曲线是下列常微分方程的解

$$\frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{u_x}{u_y} \tag{9}$$

在此例中，(9)式可用(7)式简化．结果是

$$\frac{\mathrm{d}y}{\mathrm{d}x} = c \tag{10}$$

方程(10)的解是 xy 平面上斜率为 c 的直线．这些特征线之一刚好经过平面上任意给定点(x_0, y_0)，它的方程是

$$y - y_0 = c(x - x_0) \tag{11}$$

这样的特征线的实用性是什么呢？假设求解的问题不是微分方程(7)本身而是带一个附属的条件，例如，

$$\begin{cases} u_x + cu_y = 0 \\ u(x,0) = f(x) \end{cases} \tag{12}$$

其中 f 是一个给定的函数．为了计算在任意点(x_0, y_0)上的解，我们首先确定经过这个点的特征曲线，然后沿着这条曲线到形如$(x, 0)$的点．在这样的点上，解是已知的，即 $u(x, 0)=f(x)$，并且沿着这条曲线，解是常数．因此，倘若$(x, 0)$是在经过(x_0, y_0)的特征曲线上，则 $u(x_0, y_0)$等于 $f(x)$的值．在我们的例子中，这些曲线是直线(如图 9-7 所示)．

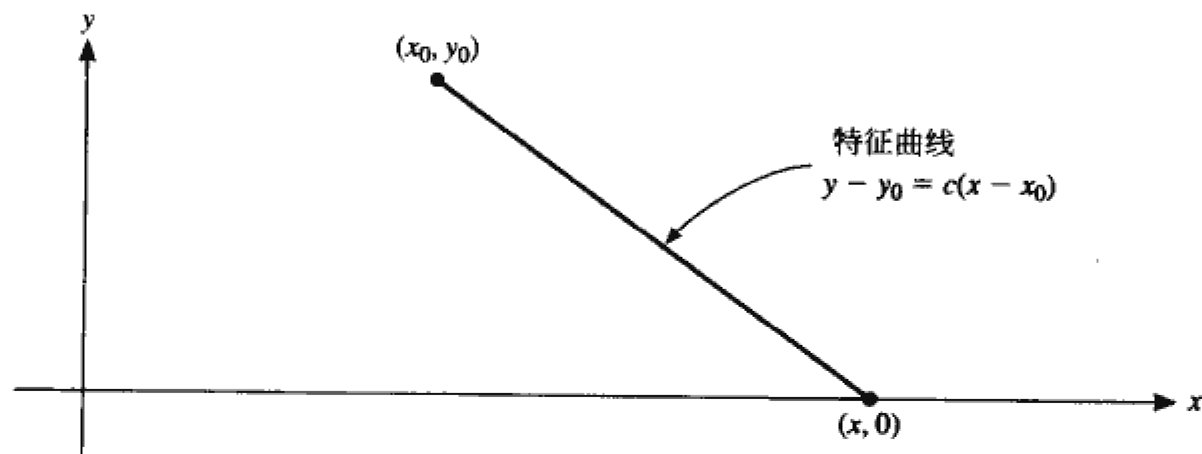


图 9-7 特征曲线：直线

[644]

$(x, 0)$位于由(11)式给出的特征曲线上的条件简化为

$$0-y_0=c(x-x_0) \tag{13}$$

于是 $x=x_0-c^{-1}y_0$ 且

$$u(x_0,y_0)=f(x_0-c^{-1}y_0) \tag{14}$$

(14)式给出原问题(12)的解．当然，我们可简单地把它写为 $u(x, y)=f(x-c^{-1}y)$．容易直接验证这个函数满足(12)式．

下面是另一个说明利用特征曲线的例子．

例 4　用特征线法解下列问题：

$$\begin{cases} u_x+yu_y=0 \\ u(0,y)=f(y) \end{cases} \tag{15}$$

解　这是一个在 xy 平面中所有点上求解的偏微分方程，而解函数的值在 y 轴上是指定的．如果我们像前面那样进行，则描述特征曲线的常微分方程是

$$\frac{\mathrm{d}y}{\mathrm{d}x}=-\frac{u_x}{u_y}=y \tag{16}$$

经过(x_0, y_0)的解是

$$y=y_0\mathrm{e}^{x-x_0} \tag{17}$$

特征曲线与 y 轴的相交出现在 $y=y_0\mathrm{e}^{-x_0}$上．于是我们有

$$u(x_0,y_0)=f(y_0\mathrm{e}^{-x_0}) \tag{18}$$

作为解．当然，通常把这个式子写为

$$u(x,y)=f(y\mathrm{e}^{-x}) \tag{19}$$

容易直接验证这是(15)式中问题的解．■

请读者尝试做一些习题——例如，习题 9.5.1，它是基本的习题．

9.5.3　特征曲线的一般理论

特征曲线的一般理论应该包括更一般的方程并且不用非对称的方式讨论变量 x 和 y．此外，我们不期望解沿特征曲线是常数；只要求它满足某个沿着特征线的常微分方程．

645

考虑方程

$$au_x+bu_y=c \tag{20}$$

其中允许 a，b 和 c 是 x，y 和 u 的函数．假如解的值在点(x_0, y_0)已知．这可以是事先计算的结果或已经指定的边界值．我们将指出如何通过积分某个常微分方程可以得到方程(20)的另外的解的值．考虑下列带有初始条件的 3 个常微分方程组：

$$\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}s}=a \qquad \dfrac{\mathrm{d}y}{\mathrm{d}s}=b \qquad \dfrac{\mathrm{d}u}{\mathrm{d}s}=c \\ x(0)=x_0 \qquad y(0)=y_0 \qquad u(0)=u_0=u(x_0,y_0) \end{cases} \tag{21}$$

为了生成一个从 $s=0$ 开始的 $x(s)$，$y(s)$和 $u(s)$的表格，我们可数值积分这个方程组．不难验证这个表格提供偏微分方程(20)的解点．实际上，

$$a(x(s),y(s),u(s))u_x(x(s),y(s))+b(x(s),y(s),u(s))u_y(x(s),y(s))$$

$$= x'(s)u_x(x(s),y(s)) + y'(s)u_y(x(s),y(s))$$

$$= \frac{\mathrm{d}}{\mathrm{d}s}u(x(s),y(s)) = c(x(s),y(s),u(s))$$

(在这个计算中，撇号表示关于 s 的微分.)

若指定 $u(x, y)$的值沿某条曲线 Γ(不是特征曲线)，则原则上我们可从 Γ 的任何点上开始积分方程组(21)，得到 $u(x, y)$沿弧的另外的值.(参见图 9-8 有助于理解这个想法.)在这个过程中得到的弧是方程(20)的特征曲线.

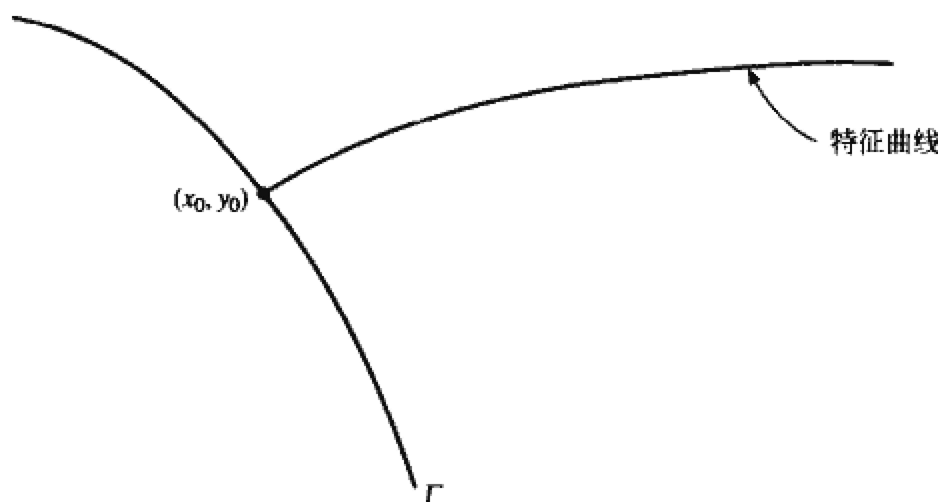


图 9-8 特征曲线：弧

例 5 确定常微分方程，使其定义属于方程

$$\sin(x^2+y^2)u_x + (3x+y^2)u_y = \mathrm{e}^{xy}$$

的特征曲线.

646

解 按照一般的理论，方程组应该是

$$\begin{cases} x' = \sin(x^2+y^2) \\ y' = 3x+y^2 \\ u' = \mathrm{e}^{xy} \end{cases} \tag{22}$$

因为在此例中 a 和 b 不含 u，所以可以只对(22)中的前两个方程积分得到特征曲线(在 xy 平面上).
■

例 6 考虑偏微分方程

$$6u_x + xu_y = y$$

给定 $u(3, 3)=4$，试问沿着特征曲线，能得到 $u(15, 21)$的怎样的值?

解 经过初始点的特征曲线由下列初始问题所决定

$$\begin{cases} x' = 6 \qquad y' = x \qquad u' = y \\ x(0) = 3 \qquad y(0) = 3 \qquad u(0) = 4 \end{cases}$$

对这些方程进行积分，得到

$$\begin{cases} x = 6s+3 \\ y = 3s^2+3s+3 \\ u = s^3 + \dfrac{3}{2}s^2 + 3s + 4 \end{cases}$$

设 $s=2$，我们得到 $x=15$，$y=21$，$u=24$. ■

例7 利用特征线法求下列边值问题的解.

$$\begin{cases} 6u_x + xu_y = y \\ u(x,y) = 4 & \text{当 } x = y \end{cases}$$

解 我们求经过点$(x, y)=(r, r)$，$r\leqslant 6$ 的特征曲线. 解方程组

$$\begin{cases} x' = 6 & y' = x & u' = y \\ x(0) = r & y(0) = r & u(0) = 4 \end{cases}$$

得到

$$\begin{cases} x = 6s + r \\ y = 3s^2 + rs + r \\ u = s^3 + \frac{1}{2}rs^2 + rs + 4 \end{cases} \tag{23}$$

647

若(x, y)是平面中已知的点，则我们试图利用(23)中前两个方程确定相应的(r, s). 计算的结果是

$$\begin{cases} r = 6 - \sqrt{(x-6)^2 + 12(x-y)} \\ s = (x-r)/6 \end{cases}$$

用 r 和 s 附近的值，从(23)中的第3个方程可算出 u 在(x, y)的值. 例如，若$(x, y)=(15, 21)$，则$(r, s)=(3, 2)$并且如例6中那样，$u=24$. ■

例8 用特征线法解边值问题

$$\begin{cases} xu_x + yuu_y = xy \\ u(x,y) = 2xy & \text{当 } xy = 3 \end{cases} \tag{24}$$

解 特征线方程是

$$\begin{cases} x' = x \\ x(0) = x_0 \end{cases} \quad \begin{cases} y' = uy \\ y(0) = 3/x_0 \end{cases} \quad \begin{cases} u' = xy \\ u(0) = 6 \end{cases}$$

因为

$$(xy)' = x'y + xy' = xy + xuy = xy(1+u) = u'(1+u)$$

所以可执行关于 s 的积分，得到

$$xy = u + \frac{1}{2}u^2 + c \tag{25}$$

利用初始条件，我们求出 $c=-21$. 可从方程(25)式解出 u，结果为

$$u = -1 + \sqrt{43 + 2xy}$$ ■

在前面的例子中，有另一种解释我们工作的方式. 我们用 r 作为参数来描述数据曲线 Γ 上的点. 如前所述，s 是一个参数，它描述在任何一条特征曲线上的点. 假定对于 Γ 上的点，$s=0$.（如图9-9所示.）用 $r=4$ 和 $s=2$ 来描述标有⊗的点.

以同样的方式，可用坐标(r, s)描述平面上的其他点. 平面上每个点都赋于 r 和 s 值未必成立. 此外，可能发生对应于两个不同的 r 值的特征曲线可能相交的情况. 此时，交点使没有

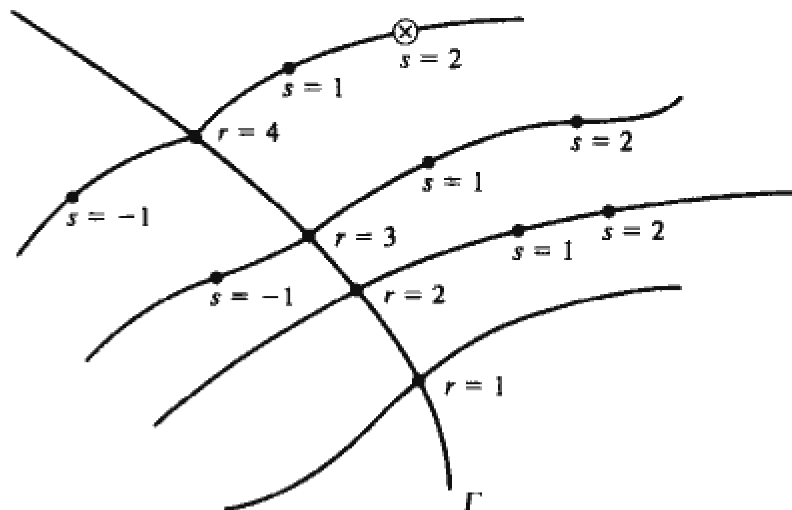


图 9-9　数据曲线 Γ

唯一的(r, s)描述．若数据曲线 Γ 不是一条特征曲线，且系数函数 a，b 和 c 是光滑的，则接近 Γ 的点将有唯一的坐标(r, s)．

648

利用刚才概述方法中的 r 和 s，可以如下描述我们的工作：用微分方程

$$\begin{cases}\dfrac{\partial x}{\partial s}=a & \dfrac{\partial y}{\partial s}=b & \dfrac{\partial u}{\partial s}=c\\ x(r,0)=f(r) & y(r,0)=g(r) & u(r,0)=h(r)\end{cases}$$

确定函数$(r, s)\mapsto x(r, s)$，$y(r, s)$和 $u(r, s)$．假定 Γ 由参数

$$x=f(r)\qquad y=g(r)$$

给出．三个微分方程的积分产生三个表示曲面的参数函数

$$x=x(r,s)\qquad y=y(r,s)\qquad u=u(r,s)$$

显然曲面包含空间曲线

$$\Gamma^*:x=f(r)\qquad y=g(r)\qquad u=h(r)$$

因此，我们已找到一个曲面，它是原偏微分方程的解并且包含一条给定的空间曲线 Γ^*．注意 Γ 是 Γ^* 到 xy 平面上的投影．

习题 9.5

1. 利用特征线法解下列问题：

a. $\begin{cases}u_x+xu_y=0\\ u(0, y)=f(y)\end{cases}$

b. $\begin{cases}u_x+2uu_y=0\\ u(0, y)=f(y)\end{cases}$

c. $\begin{cases}xu_x+2yu_y=0\\ u(1, y)=f(y)\end{cases}$

649

2. 给出

$$\begin{cases}u_x+yu_y=0\\ u(18,3e)=k\pi/2\end{cases}$$

求值 $u(17, 3)$．提示：可利用课本中的例子．

3. 验证例8解答中得到的函数是给定问题的解.

4. 求例6中微分方程的解. 已知 $u=e^x\sin y$ 在曲线 $y=x^3$ 上. 说明如何克服数值上的困难. 利用点(7, 5)在经过(1, 1)的特征曲线上的这一事实，求 $u(7, 5)$.

5. 按照一般的理论，说明为什么方程(7)和(15)的解函数沿着特征曲线是常数.

6. 说明例7中求得的解不唯一. 提示：当(x, y)给定时，在求解(r, s)的过程中，二次方程有两个解.

7. 用特征线法求解

$$\begin{cases}u_x+u_y=u^2\\ u(x,y)=y \quad \text{在直线 } x+y=0 \text{ 上}\end{cases}$$

8. 用特征线法解下列边值问题并说明所有困难：

$$\begin{cases}u_x+2u_y=u\\ u=1 \quad \text{当 } y=2x\end{cases}$$

9. 用特征线法求解

$$\begin{cases}uu_x+u_y=1\\ u=r \quad \text{在曲线 } x=r^2, y=2r \text{ 上}\end{cases}$$

10. 用特征线法求解

$$\begin{cases}u_x+2u_y=y\\ u=r^2 \quad \text{在圆 } x=\cos r, y=\sin r \text{ 上}\end{cases}$$

11. 考虑偏微分方程 $au_x+bu_y=0$，其中 a 和 b 仅仅是 x 的函数. 求经过(x_0, y_0)的曲线方程，在此曲线上 u 的值保持不变.

9.6 拟线性二阶方程：特征线法

在本节中，讨论方程

$$au_{xx}+bu_{xy}+cu_{yy}+e=0 \tag{1}$$

我们允许 a，b，c 和 e 是 x，y，u，u_x 和 u_y 的函数. 这样的一个方程称为**拟线性的**. 正如在对一阶方程的特征曲线的研究中那样，这里我们询求解函数沿着 xy 平面上的一条曲线有怎样的特性.

650

9.6.1 特征曲线

设 C 是 xy 平面上用参数

$$x=x(s) \qquad y=y(s) \qquad (s\in\mathbb{R})$$

给出的一条曲线. 传统的记号 $p=u_x$ 和 $q=u_y$ 将简化我们的工作. x，y，p，q 和 u 不直接作为 s 的函数. 通过直接微分，有

$$\frac{\mathrm{d}p}{\mathrm{d}s}=\frac{\partial p}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}s}+\frac{\partial p}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}s}=u_{xx}x'+u_{xy}y' \tag{2}$$

$$\frac{\mathrm{d}q}{\mathrm{d}s}=\frac{\partial q}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}s}+\frac{\partial q}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}s}=u_{xy}x'+u_{yy}y' \tag{3}$$

从(2)式可解 u_{xx}，得到

$$u_{xx}=\left(\frac{\mathrm{d}p}{\mathrm{d}s}-u_{xy}y'\right)/x'=\frac{\mathrm{d}p}{\mathrm{d}s}\frac{\mathrm{d}s}{\mathrm{d}x}-u_{xy}\frac{\mathrm{d}y}{\mathrm{d}s}\frac{\mathrm{d}s}{\mathrm{d}x}=\frac{\mathrm{d}p}{\mathrm{d}x}-u_{xy}\frac{\mathrm{d}y}{\mathrm{d}x} \tag{4}$$

类似地，从(3)式有

$$u_{yy}=\left(\frac{\mathrm{d}q}{\mathrm{d}s}-u_{xy}x'\right)/y'=\frac{\mathrm{d}q}{\mathrm{d}s}\frac{\mathrm{d}s}{\mathrm{d}y}-u_{xy}\frac{\mathrm{d}x}{\mathrm{d}s}\frac{\mathrm{d}s}{\mathrm{d}y}=\frac{\mathrm{d}q}{\mathrm{d}y}-u_{xy}\frac{\mathrm{d}x}{\mathrm{d}y} \tag{5}$$

(4)和(5)这些方程沿着曲线 C 是成立的. 如果把刚才导出的表达式代入到方程(1)，结果是

$$a\left(\frac{\mathrm{d}p}{\mathrm{d}x}-u_{xy}\frac{\mathrm{d}y}{\mathrm{d}x}\right)+bu_{xy}+c\left(\frac{\mathrm{d}q}{\mathrm{d}y}-u_{xy}\frac{\mathrm{d}x}{\mathrm{d}y}\right)+e=0 \tag{6}$$

当(6)式用 $\mathrm{d}y/\mathrm{d}x$ 相乘时，结果为

$$a\left[\frac{\mathrm{d}p}{\mathrm{d}x}\frac{\mathrm{d}y}{\mathrm{d}x}-u_{xy}\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2\right]+bu_{xy}\frac{\mathrm{d}y}{\mathrm{d}x}+c\left[\frac{\mathrm{d}q}{\mathrm{d}x}-u_{xy}\right]+e\frac{\mathrm{d}y}{\mathrm{d}x}=0 \tag{7}$$

合并(7)中的项，有

$$-u_{xy}\left[a\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2-b\frac{\mathrm{d}y}{\mathrm{d}x}+c\right]+a\frac{\mathrm{d}p}{\mathrm{d}x}\frac{\mathrm{d}y}{\mathrm{d}x}+c\frac{\mathrm{d}q}{\mathrm{d}x}+e\frac{\mathrm{d}y}{\mathrm{d}x}=0 \tag{8}$$

选择直到现在还未指定的曲线 C 使得(8)式中不出现项 u_{xy}. 因而，C 被描述成微分方程

$$a\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2-b\frac{\mathrm{d}y}{\mathrm{d}x}+c=0 \tag{9}$$

这样的一条曲线 C 被称为微分方程(1)的**特征曲线**.

651

9.6.2 分类

因为(9)式是 $\mathrm{d}y/\mathrm{d}x$ 的二次方程，所以特征曲线的性态由**判别式**

$$\Delta\equiv b^2-4ac$$

确定. 若在某个值(x, y, u)上 $\Delta>0$，则称微分方程是**双曲型的**；若 $\Delta=0$，则称微分方程是**抛物型的**；若 $\Delta<0$，则称它是**椭圆型的**. 这个分类可随 xy 平面内的点而变化，它也与解 u 有关，因为 a，b 和 c 允许与 x，y，u，u_x 和 u_y 有关. 在**线性情况**中，系数函数 a，b 和 c 只与 x 和 y 有关，因而分类变得更简单.

用三个代表性的熟悉方程很好地说明了这个方程的分类.

名称	形式	a	b	c	Δ	类型
热传导	$u_{xx}-u_y=0$	1	0	0	0	抛物型
波动	$u_{xx}-u_{yy}=0$	1	0	-1	4	双曲型
拉普拉斯	$u_{xx}+u_{yy}=0$	1	0	1	-4	椭圆型

例 1 对下列方程进行分类

$$(x+y)u_{xx}+(1+x^2)u_{yy}=0$$

解 这个方程的判别式是

$$\Delta(x,y)=-4(x+y)(1+x^2)$$

在 $x+y>0$ 的区域中，方程是椭圆型的. 在 $x+y<0$ 的区域中，方程是双曲型的，而在直线 $x+y=0$ 上，它是抛物型的. ■

例 2 求属于下列方程的特征曲线

$$yu_{xx}+(x+y^2)u_{xy}+xyu_{yy}=0 \tag{10}$$

解 描述特征曲线的微分方程是(9)式. 在此例中，它是

$$y\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2-(x+y^2)\frac{\mathrm{d}y}{\mathrm{d}x}+xy=0 \tag{11}$$

判别式是

$$\Delta=b^2-4ac=(x+y^2)^2-4xy^2=(x-y^2)^2$$

因为 $\Delta>0$(除了在曲线 $x=y^2$ 上之外)，偏微分方程(10)是双曲型的，所以特征曲线有两个常微分方程求解，即

$$\frac{\mathrm{d}y}{\mathrm{d}x}=\frac{b\pm\sqrt{\Delta}}{2a}=\frac{(x+y^2)\pm|x-y^2|}{2y} \tag{12}$$

652 假设 $x>y^2$. 对应于加号的第一个方程是

$$\frac{\mathrm{d}y}{\mathrm{d}x}=\frac{x}{y}$$

其解是双曲线族 $y^2-x^2=\alpha$. 对应于减号的第二个方程是

$$\frac{\mathrm{d}y}{\mathrm{d}x}=y$$

其解是指数曲线族 $y=\beta e^x$. 当 $x<y^2$ 时，这两种情况相反. 当 $x=y^2$ 时，方程(10)是抛物型的且特征曲线再次是常微分方程 $\mathrm{d}y/\mathrm{d}x=y$ 的解. ■

9.6.3 算法

现在我们回到(8)式并假定曲线 C 是特征曲线. 在 C 上，斜率函数 $\mathrm{d}y/\mathrm{d}x$ 服从(9)式，故(8)式简化为

$$a\frac{\mathrm{d}p}{\mathrm{d}x}\frac{\mathrm{d}y}{\mathrm{d}x}+c\frac{\mathrm{d}q}{\mathrm{d}x}+e\frac{\mathrm{d}y}{\mathrm{d}x}=0 \tag{13}$$

现在我们指出，在双曲型方程情况中，如何局部地使用(13)式来数值计算一个解. 此时，两条特征曲线穿过一个已知点(x_0, y_0). 考虑图 9-10 中指出的情况，其中两条特征曲线穿过点 A.

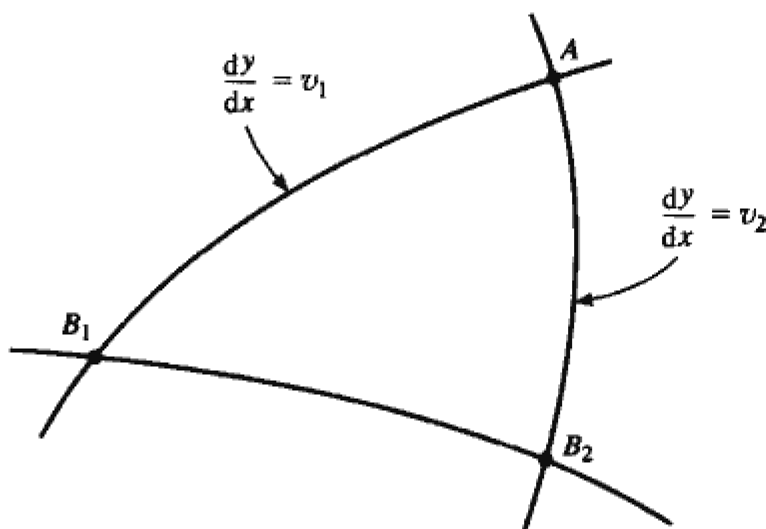


图 9-10 特征曲线 AB_1 和 AB_2

为简化记号，我们引进解二次方程(9)中产生的函数 v_1 和 v_2：

653

$$\begin{cases}v_1=(b+\sqrt{\Delta})/(2a)\\ v_2=(b-\sqrt{\Delta})/(2a)\end{cases} \tag{14}$$

因而两条特征曲线由下列微分方程给出

$$\frac{dy}{dx}=v_1 \qquad \frac{dy}{dx}=v_2 \tag{15}$$

(13)式可以用下列事实简化：(14)式中的函数 v_1 和 v_2 满足关系：

$$v_1+v_2=b/a \qquad v_1v_2=c/a \tag{16}$$

对两条特征曲线，(13)式的更简单的形式是

$$\begin{cases}\dfrac{dp}{dx}+v_1\dfrac{dq}{dx}=-e/a & 当\dfrac{dy}{dx}=v_2\\[2ex] \dfrac{dp}{dx}+v_2\dfrac{dq}{dx}=-e/a & 当\dfrac{dy}{dx}=v_1\end{cases} \tag{17}$$

可使用有限差分法求解(17). 求解的方程是方程(15)的离散化形式：

$$\frac{y(A)-y(B_1)}{x(A)-x(B_1)}=\frac{v_1(A)+v_1(B_1)}{2} \tag{18}$$

$$\frac{y(A)-y(B_2)}{x(A)-x(B_2)}=\frac{v_2(A)+v_2(B_2)}{2} \tag{19}$$

和方程(17)的离散化形式：

$$\frac{p(A)-p(B_2)}{x(A)-x(B_2)}+\left[\frac{v_1(A)+v_1(B_2)}{2}\right]\left[\frac{q(A)-q(B_2)}{x(A)-x(B_2)}\right]$$
$$=-\frac{1}{2}[(e/a)(A)+(e/a)(B_2)] \tag{20}$$

$$\frac{p(A)-p(B_1)}{x(A)-x(B_1)}+\left[\frac{v_2(A)+v_2(B_1)}{2}\right]\left[\frac{q(A)-q(B_1)}{x(A)-x(B_1)}\right]$$
$$=-\frac{1}{2}[(e/a)(A)+(e/a)(B_1)] \tag{21}$$

这里 A 和 B_1 是特征曲线 $dy/dx=v_1$ 上的点，而 A 和 B_2 是特征曲线 $dy/dx=v_2$ 上的点. 计算 $u(A)$ 的公式是

$$u(A)=u(B_1)+\left[\frac{p(A)+p(B_1)}{2}\right][x(A)-x(B_1)]$$
$$+\left[\frac{q(A)+q(B_1)}{2}\right][y(A)-y(B_1)] \tag{22}$$

这是等式

$$du=u_x dx+u_y dy=p dx+q dy$$

[654]

的有限差分模拟，因为 $p=u_x$ 且 $q=u_y$. 注意到，在(18)～(22)式中系统地使用了沿不同弧的函数的平均值.

现在假定我们知道 x，y，u，p 和 q 在两个点 B_1 和 B_2 上的值，则(18)～(22)式可看作是确定新值 $x(A)$，$y(A)$，$u(A)$，$p(A)$ 和 $q(A)$ 的方程. 求新值问题本身会引起困难，因为这些方程是非线性的. 在数值实践中，这些方程通常用迭代法求解. 下面是这一个过程的概要：

1. 从 $x(A)$，$y(A)$，$u(A)$，$p(A)$ 和 $q(A)$ 的正确值的猜测开始. 这些猜测可能是这些函数最新的值(譬如说在 B_1)的简单扰动.

2. 利用 $x(A)$，$y(A)$，$p(A)$，$q(A)$和 $u(A)$最新的值计算 $v_1(A)$，$v_2(A)$和$(e/a)(A)$.

3. 利用(18)和(19)式再计算 $y(A)$和 $x(A)$. 然后利用(20)和(21)式再计算 $p(A)$和 $q(A)$. 利用(22)式再计算 $u(A)$. 如果新值与旧值差别很大，则退回第 2 步.

在第 1 步，利用(18)～(21)式的比较粗糙的形式，容易得到计算值的初始猜测. 因此，可利用在 B_1 和 B_2 上的值替代在这些方程中出现的平均值. 当这些做完后，方程(18)和(19)是线性的并且很快地解出迭代的初值. 这个工作产生下列公式：

$$x(A) = \frac{y(B_2) - y(B_1) + x(B_1)v_1(B_1) - x(B_2)v_2(B_2)}{v_1(B_1) - v_2(B_2)}$$

$$y(A) = y(B_1) + v_1(B_1)[x(A) - x(B_1)]$$

$$R = p(B_2) - p(B_1) + v_1(B_2)q(B_2) - v_2(B_1)q(B_1)$$

$$S = (e/a)(B_2)[x(A) - x(B_2)] - (e/a)(B_1)[x(A) - x(B_1)]$$

$$q(A) = (R - S)/[v_1(B_2) - v_2(B_1)]$$

$$p(A) = p(B_2) - (e/a)(B_2)[x(A) - x(B_2)] - v_1(B_2)[q(A) - q(B_2)]$$

若拟线性方程(1)事实上是线性的，则 a，b，c，Δ，v_1和 v_2 仅仅是 x 和 y 的函数. 函数 u 及其导数将不出现前面的 6 个函数中. 此时(18)和(19)式可和 $x(A)$和 $y(A)$一起求解. 之后，(20)和(21)式可和 $p(A)$和 $q(A)$一起求解.

例 3 编写用特征线法求解边值问题

$$\begin{cases} u_{xx} - 4u_{yy} - u_y = 0 \\ u(x,0) = f(x) \quad u_y(x,0) = g(x) \qquad (0 \leqslant x \leqslant 1) \end{cases}$$

的简略的伪代码.

655

解 我们在区间[0, 1]中选取 n 个等距点 x_j，计算起点在 x_j 的特征曲线交点上的数值解. 因为 $\Delta=16$，所以(14)式给出 $v_1=2$ 和 $v_2=-2$. 特征曲线的微分方程是 $\mathrm{d}y/\mathrm{d}x=2$ 和 $\mathrm{d}y/\mathrm{d}x=-2$. 如图 9-11 所示，这些曲线是直线，其中取 $n=8$. 在此例中，因为 $q=u_y=-e$，所以方程(17)具有下列形式：

$$\begin{cases} \dfrac{\mathrm{d}p}{\mathrm{d}x} + 2\dfrac{\mathrm{d}q}{\mathrm{d}x} - q = 0 \qquad 当\dfrac{\mathrm{d}y}{\mathrm{d}x} = -2 \\ \dfrac{\mathrm{d}p}{\mathrm{d}x} - 2\dfrac{\mathrm{d}q}{\mathrm{d}x} - q = 0 \qquad 当\dfrac{\mathrm{d}y}{\mathrm{d}x} = 2 \end{cases}$$

现设 $h=1/(n-1)$. 这是区间[0, 1]上离散点 x_j 之间的距离. 参考图 9-12，我们看到若 $B_1=(x, y)$，则 $B_2=(x+h, y)$，$A=(x+\frac{h}{2}, y+h)$. 因此，可省略(18)和(19)式. 在(20)式和(21)式中使用 $e=-u_y=-q$. 化简以后，这些式子变成

$$\begin{cases} p(A) + \left(2 + \dfrac{h}{4}\right)q(A) = p(B_2) + \left(2 - \dfrac{h}{4}\right)q(B_2) \\ p(A) - \left(2 + \dfrac{h}{4}\right)q(A) = p(B_1) - \left(2 - \dfrac{h}{4}\right)q(B_1) \end{cases}$$

这对线性方程的解是

$$p(A) = \frac{1}{2}[p(B_1) + p(B_2)] + \left(1 - \frac{h}{8}\right)[q(B_2) - q(B_1)]$$

$$q(A) = \left\{p(B_2) - p(B_1) + \left(2 - \frac{h}{4}\right)[q(B_2) + q(B_1)]\right\}\Big/\left(4 + \frac{h}{2}\right)$$

适当化简后，公式(22)化为

$$u(A) = u(B_1) + \frac{h}{4}[p(A) + p(B_1)] + \frac{h}{2}[q(A) + q(B_1)]$$

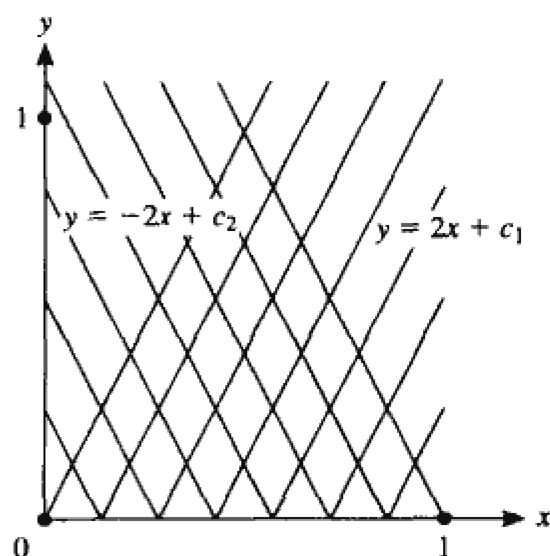


图 9-11 例 3 中的特征曲线

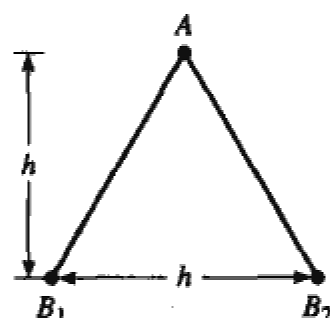


图 9-12 例 3 中离散部分空间的说明

下面是这个过程的算法：

```
input n
h←1/(n−1)
for j=1 to n do
    x_j←(j−1)h
    y_j←0
    u_j←f(x_j)
    p_j←f'(x_j)
    q_j←g(x_j)
    output x_j, y_j, p_j, q_j, u_j
end do
for i=1 to n do
    y_i←ih
    for j=1 to n−i do
        x_j←x_j+h/2
        p̃←[p_j+p_{j+1}]/2+(1−h/8)[q_{j+1}−q_j]
        q̃←[p_{j+1}−p_j+(2−h/4)(q_{j+1}+q_j)]/(4+h/2)
        u_j←u_j+h(p̃+p_j)/4+h(q̃+q_j)/2
        p_j←p̃
        q_j←q̃
        output x_j, y_j, p_j, q_j, u_j
    end do
end do
```

因此，我们已经对给出的边界问题讨论了一个简略的代码. ■

9.6.4 另一种特征线法

有另一种特征线理论的方法. 我们首先分析变量从(x, y)到(ξ, η)的一个变化能导致微分方程(1)中什么简化. 假设引进新变量

$$\xi = \xi(x,y) \qquad \eta = \eta(x,y) \tag{23}$$

656~657

这个替换在微分方程

$$au_{xx} + bu_{xy} + cu_{yy} + e = 0 \tag{24}$$

中的影响可首先通过计算

$$\begin{aligned}
u_x &= u_\xi \xi_x + u_\eta \eta_x \\
u_y &= u_\xi \xi_y + u_\eta \eta_y \\
u_{xx} &= u_\xi \xi_{xx} + u_{\xi\xi}\xi_x^2 + u_{\xi\eta}\eta_x\xi_x + u_\eta\eta_{xx} + u_{\eta\xi}\xi_x\eta_x + u_{\eta\eta}\eta_x^2 \\
u_{xy} &= u_\xi \xi_{xy} + u_{\xi\xi}\xi_y\xi_x + u_{\xi\eta}\eta_y\xi_x + u_\eta\eta_{xy} + u_{\eta\xi}\xi_y\eta_x + u_{\eta\eta}\eta_y\eta_x \\
u_{yy} &= u_\xi \xi_{yy} + u_{\xi\xi}\xi_y^2 + u_{\xi\eta}\eta_y\xi_y + u_\eta\eta_{yy} + u_{\eta\xi}\xi_y\eta_y + u_{\eta\eta}\eta_y^2
\end{aligned}$$

来确定. 把这些表达式代入(24)式，我们得到

$$\begin{aligned}
u_{\xi\xi}[a\xi_x^2 + b\xi_x\xi_y + c\xi_y^2] + u_{\xi\eta}[2a\xi_x\eta_x + b\xi_x\eta_y + b\xi_y\eta_x + 2c\xi_y\eta_y] \\
+ u_{\eta\eta}[a\eta_x^2 + b\eta_y\eta_x + c\eta_y^2] + f = 0
\end{aligned} \tag{25}$$

在 a, b, c 和 e 中，必须也作适当的替代. 所有二阶导数项都被显示，而 f 包含所有其他项. (见习题 9.6.8.) 因为我们对双曲型情况有兴趣，所以二次方程

$$a\lambda^2 + b\lambda + c = 0$$

应假定有两个实根$(-b \pm \sqrt{\Delta})/(2a)$，其中判别式 $\Delta = b^2 - 4ac$ 是正的. 若我们选择坐标变换使得

$$\xi_x/\xi_y = (-b + \sqrt{\Delta})/(2a)$$
$$\eta_x/\eta_y = (-b - \sqrt{\Delta})/(2a)$$

则微分方程(25)化简为

$$u_{\xi\eta}[2a\xi_x\eta_x + b(\xi_x\eta_y + \xi_y\eta_x) + 2c\xi_y\eta_y] + f = 0 \tag{26}$$

这个方程是双曲型方程的**典范型**. 用

$$\xi(x,y) = \text{常数} \qquad \eta(x,y) = \text{常数} \tag{27}$$

描述的曲线是**特征曲线**. 容易验证这些曲线是前面(19)式中得到的微分方程的解.

例 4 求使方程

$$u_{xx} - yu_{yy} = 0 \tag{28}$$

是双曲型的区域，并确定化方程为典范型的变量替换.

解 因为判别式是 $b^2 - 4ac = 4y$，所以方程(28)在上半平面内是双曲型的. 为求变量替换，我们解微分方程

658

$$\xi_x/\xi_y = (-b + \sqrt{b^2 - 4ac})/(2a) = y^{1/2}$$
$$\eta_x/\eta_y = (-b - \sqrt{b^2 - 4ac})/(2a) = -y^{1/2}$$

作为解，我们可取

$$\xi = x + 2y^{1/2} \qquad \eta = x - 2y^{1/2}$$

根据新变量，方程(28)变成

$$4u_{\xi\eta} + \frac{2}{\xi - \eta}(u_\xi - u_\eta) = 0$$

■

习题 9.6

1. 证明：若 $b=0$ 且 $ac<0$，则微分方程(1)是双曲型的. 说明它的特征曲线是 $\mathrm{d}y/\mathrm{d}x=\sqrt{-c/a}$的解.
2. 证明：若 $c=-a$ 且 $b^2>4a^2$，则微分方程(1)是双曲型的，且穿过 xy 平面上每个点的两条特征曲线是相互垂直的.
3. 求特征曲线是 $x\cos\alpha+y\sin\alpha=0$ 和 $x^2+y^2=\beta$ 的二阶偏微分方程.
4. 将下列方程分类成双曲型的，抛物型的或椭圆型的：
   a. $yu_{xx}+xu_{xy}+u_{yy}+u+u_x=0$
   b. $xyu_{xy}+e^x u_x+yu_y=0$
   c. $3u_{xx}+u_{xy}+u_{yy}+2yu+7=0$
5. 求方程 $xy^2u_{xx}=u_{yy}$ 是双曲型的区域. 确定它的典范型及简化的变量替换.
6. 验证沿 $\xi(x, y)$或 $\eta(x, y)$保持常数的曲线是方程(9)中定义的特征曲线.
7. 作(23)式中的变量替换后，(24)式变成(25)式，其具有形式

   $$\alpha u_{\xi\xi} + \beta u_{\xi\eta} + \gamma u_{\eta\eta} + e = 0$$

   证明

   $$\beta^2 - 4\alpha\gamma = (b^2 - 4ac)J^2$$

   其中 J 是变换的雅可比行列式

   $$J = \begin{vmatrix} \xi_x & \xi_y \\ \eta_x & \eta_y \end{vmatrix}$$

   最后，得到结论：若雅可比行列式不为零，则微分方程的类型(椭圆型的，抛物型的或双曲型的)不改变.
8. 在(25)式中，证明

   $$f = e + au_\xi\xi_{xx} + au_\eta\eta_{xx} + bu_\xi\xi_{xy} + bu_\eta\eta_{xy} + cu_\xi\xi_{yy} + cu_\eta\eta_{yy}$$

9. 证明例 4 中的微分方程的典范型是

   $$16u_{\xi\eta} + u_\xi + u_\eta = 0$$

10. 补充前面例 3 中公式推导的细节
11. 计算出例 3 中给出的解.
12. 验证例 4 中最后的方程.

659

计算机习题 9.6

利用特征线方法，编写解下列双曲型问题的通用代码

$$\begin{cases} au_{xx} + bu_{yy} + cu_x + du_y = 0 \\ u(0,y) = f(y) \quad u_x(0,y) = g(y) \quad (0 \leqslant y \leqslant 1) \end{cases}$$

假定 a，b，c 和 d 是满足 $ab<0$ 的常数. 在 y 轴上区间 $0\leqslant y\leqslant 1$ 内取等距点.

9.7 双曲型问题的其他方法

我们将首先讨论双曲型一阶线性偏微分方程组的一些有限差分法. 从伴随初始条件的单个方程入手：

$$\begin{cases} u_t = \alpha u_x \\ u(x,0) = f(x) \qquad (-\infty < x < \infty) \end{cases} \tag{1}$$

这里 α 是(实)常数. 因为在直线 $t=0$ 上的解是已知的，所以尝试行**进方法**延伸解到另外的直线 $t=k$，$t=2k$，等等是自然的，其中 k 是 t 变量采用的步长.

9.7.1 拉克斯-温德罗夫方法

假设解 u 关于两个变量 x 和 t 是任意次连续可微的，则由泰勒定理，得

$$u(x,t+k) = u + ku_t + \frac{k^2}{2!}u_{tt} + \frac{k^3}{3!}u_{ttt} + \cdots \tag{2}$$

在这个等式的右边，u 及其导数在基点(x, t)上求值. 另外，级数应该在某项终止并且补充一个适当的误差项.

因为函数 u 是(1)的解，所以 $u_t = \alpha u_x$，$u_{tt} = \alpha^2 u_{xx}$ 等等.（习题 9.7.1 要求对这给出证明.）因此，(2)式可写成

$$u(x,t+k) = u + k\alpha u_x + \frac{(k\alpha)^2}{2!}u_{xx} + \frac{(k\alpha)^3}{3!}u_{xxx} + \cdots \tag{3}$$

当我们希望设计一个具有步长的二阶精度的数值方法时，(3)中的级数在项 u_{xx} 处截断. 然后，(3)式右边的导数可用二阶有限差分近似代替. 设 h 是 x 变量采取的步长. 则步进求解的公式是

660

$$v(x,t+k) = v(x,t) + k\alpha\left[\frac{v(x+h,t) - v(x-h,t)}{2h}\right] + \frac{k^2\alpha^2}{2}\left[\frac{v(x+h,t) - 2v(x,t) + v(x-h,t)}{h^2}\right]$$

使得

$$v(x,t+k) = (2s^2 + s)v(x+h,t) + (1-4s^2)v(x,t) + (2s^2 - s)v(x-h,t) \tag{4}$$

其中 $s = (\alpha k)/(2h)$. (4)式表达的方法是**拉克斯-温德罗夫方法**. 因为方程(4)的解一般不是方程(3)的解，所以我们改变成变量 v.

显然，不存在从方程(3)中导出高阶方法的限制. 习题 9.7.2 要求利用这些原则导出一个三阶方法.

稳定性分析

在由(1)式给出的简单的模型问题中，立即可写出解

$$u(x,t) = f(x + \alpha t)$$

若引入网格点(jh, nk)，并且取 $v_{jn} = v(jh, nk)$，则(4)式将具有下列形式：

$$v_{j,n+1} = (2s^2 + s)v_{j+1,n} + (1-4s^2)v_{jn} + (2s^2 - s)v_{j-1,n} \tag{5}$$

其中 $s = (k\alpha)/(2h)$. 利用 9.1 节的傅里叶方法可执行稳定性分析. 为此，我们寻找(5)式具有下列形式的解

$$v_{jn} = e^{ij\beta h}e^{n\lambda k} \qquad (i = \sqrt{-1})$$

代入这个试验解到(5)式中并作化简，得到

$$e^{\lambda k} = 1 - 4s^2 + e^{i\beta h}(2s^2 + s) + e^{-i\beta h}(2s^2 - s)$$

$$= 1 - 4s^2 + s(e^{i\beta h} - e^{-i\beta h}) + 2s^2(e^{i\beta h} + e^{-i\beta h})$$
$$= 1 - 4s^2 + 2is\sin\beta h + 4s^2\cos\beta h$$

这里使用了欧拉关系

$$e^{i\beta h} = \cos\beta h + i\ \sin\beta h$$

为了稳定性，我们要求 $|e^{\lambda h}| \leqslant 1$. 计算得到

$$|e^{\lambda h}|^2 = 1 - 16s^2\sin^2\theta(1 - 4s^2\sin^2\theta + \cos^2\theta) \tag{6}$$

661

其中 $\theta=\beta h/2$. 现在稳定性条件是

$$1 - 4s^2\sin^2\theta + \cos^2\theta \geqslant 0 \tag{7}$$

当 $\theta=\pi/4$ 时，左边表达式出现极小值，当 $|s| \leqslant 1/2$ 时，极小值是非负的. 因此稳定性要求(必要条件)是 $k|\alpha| \leqslant h$.

9.7.2 方程组

下面考虑双曲型方程组，它具有下列形式

$$U_t = AU_x \tag{8}$$

这里 U 是一个分量为函数 $u^{(1)}$，$u^{(2)}$，…，$u^{(m)}$ 的向量. 每个分量都是 (x, t) 的函数. 矩阵 A 是 $m\times m$ 的. 如果问题是双曲型的，则 A 必有 m 个不同的实特征值. 拉克斯-温德罗夫方法的向量形式是与(4)式类似的：

$$V(x,t+k) = V(x,t) + \left(\frac{k}{2h}\right)A[V(x+h,t) - V(x-h,t)]$$
$$+ \left(\frac{k^2}{2h^2}\right)A^2[V(x+h,t) - 2V(x,t) + V(x-h,t)] \tag{9}$$

使得

$$V(x,t+h) = (\tau A)(2\tau A + I)V(x+h,t) + (I - (2\tau A)^2)V(x,t)$$
$$+ (\tau A)(2\tau A - I)V(x-h,t)$$

其中 $\tau=k/(2h)$. 不作详细讨论，我们只提出这个数值方法是稳定的当且仅当 A 的每个特征值 λ 满足 $k|\lambda| \leqslant h$，即 A 的谱半径满足不等式 $k\rho(A) \leqslant h$.

9.7.3 温德罗夫隐式方法

前面的讨论应用于只提供初值的微分方程. 如果变量 x 限定于一个区间，例如[0, 1]，则一个适定问题会指定区域 $0\leqslant x\leqslant 1$ 和 $t\geqslant 0$ 的整个边界值. 此时，隐式数值方法可用于微分方程(1). 根据 9.2 节的讨论可以期望隐式方法可能得到较好的稳定性质. 一个这样的方法是由温德罗夫(Wendroff)名字命名的. 首先写出它原来提出的形式：

$$\frac{1}{2}\left(\frac{v_{j,n+1} - v_{j,n}}{k} + \frac{v_{j+1,n+1} - v_{j+1,n}}{k}\right)$$
$$= \frac{\alpha}{2}\left(\frac{v_{j+1,n} - v_{jn}}{h} + \frac{v_{j+1,n+1} - v_{j,n+1}}{h}\right) \tag{10}$$

这里可以看出用一阶差分的平均值来表示导数. 这个方法的截断误差是 $\mathcal{O}(h^2+k^2)$.

662

稳定性分析

在进行稳定性分析之前，我们把(10)式写成下列形式：

$$v_{j,n+1}(1+r)+v_{j+1,n+1}(1-r)=v_{jn}(1-r)+v_{j+1,n}(1+r) \tag{11}$$

其中 $r=\alpha k/h$. 下面，我们寻找形式为

$$v_{jn}=e^{ij\beta h}e^{n\lambda k}$$

的方程(11)的解. 把它代入方程(11)，接着化简它得到

$$e^{\lambda k}(1+r)+e^{\lambda k}e^{i\beta h}(1-r)=1-r+e^{i\beta h}(1+r)$$

因此，

$$e^{\lambda k}=[1+e^{i\beta h}+r(e^{i\beta h}-1)]/[1+e^{i\beta h}-r(e^{i\beta h}-1)] \tag{12}$$

为使 $|e^{\lambda k}|\leqslant 1$，其充要条件是

$$|1+e^{i\beta h}+r(e^{i\beta h}-1)|\leqslant|1+e^{i\beta h}-r(e^{i\beta h}-1)| \tag{13}$$

求助于习题 9.7.6，我们找到一个等价条件

$$(1+\cos\beta h)(r\cos\beta h-r)+(\sin\beta h)(r\sin\beta h)\leqslant 0 \tag{14}$$

这个式子的左边为 0，故方法对一切 α，k 和 h 是稳定的.

误差分析

现在我们着手证明前面作出的断言：温德罗夫方法中的截断误差是 $\mathcal{O}(h^2+k^2)$. 其确切含义需要某些附加的说明. 微分方程具有形式 $Lu=0$，其中 L 定义为 $Lu\equiv u_t-\alpha u_x$. 由(10)式，在数值过程中替代 L 的有限差分算子具有形式

$$\frac{1}{2}(A+B)-\frac{1}{2}\alpha(C+D)$$

其中 A，B，C 和 D 由下式给出

$$(Au)(x,t)=k^{-1}[u(x,t+k)-u(x,t)]$$
$$(Bu)(x,t)=k^{-1}[u(x+h,t+k)-u(x+h,t)]$$
$$(Cu)(x,t)=h^{-1}[u(x+h,t)-u(x,t)]$$
$$(Du)(x,t)=h^{-1}[u(x+h,t+k)-u(x,t+k)]$$

利用中心差分，我们有

$$(Au)(x,t)=u_t\left(x,t+\frac{1}{2}k\right)+\mathcal{O}(k^2)$$

$$(Bu)(x,t)=u_t\left(x+h,t+\frac{1}{2}k\right)+\mathcal{O}(k^2)$$

$$(Cu)(x,t)=u_x\left(x+\frac{1}{2}h,t\right)+\mathcal{O}(h^2)$$

663

$$(Du)(x,t)=u_x\left(x+\frac{1}{2}h,t+k\right)+\mathcal{O}(h^2)$$

对任意充分光滑的 u，我们不坚持

$$\left\{L-\frac{1}{2}(A+B)+\frac{1}{2}\alpha(C+D)\right\}u=\mathcal{O}(h^2+k^2)$$

仅对满足微分方程的函数证明这个等式. 因此在证明中，假设 $Lu=0$，从而我们只需证明

$$(A+B)u-\alpha(C+D)u=\mathcal{O}(h^2+k^2)$$

引理 1(离散化误差引理) 温德罗夫方法中的离散化误差是 $\mathcal{O}(h^2+k^2)$.

证明 由泰勒定理和微分方程，我们有

$$A = u_t(x,t) + \frac{1}{2}ku_{tt}(x,t) + \mathcal{O}(k^2)$$
$$= \alpha u_x(x,t) + \frac{1}{2}\alpha^2 ku_{xx}(x,t) + \mathcal{O}(k^2)$$

在后面等式中用 $x+h$ 代替 x 得到

$$B = \alpha u_x(x+h,t) + \frac{1}{2}\alpha^2 ku_{xx}(x+h,t) + \mathcal{O}(k^2)$$
$$= \alpha u_x(x,t) + \alpha hu_{xx}(x,t) + \mathcal{O}(h^2) + \frac{1}{2}\alpha^2 ku_{xx}(x,t) + \mathcal{O}(hk) + \mathcal{O}(k^2)$$

类似地，我们得到

$$C = u_x(x,t) + \frac{1}{2}hu_{xx}(x,t) + \mathcal{O}(h^2)$$
$$D = u_x(x,t+k) + \frac{1}{2}hu_{xx}(x,t+k) + \mathcal{O}(h^2)$$
$$= u_x(x,t) + ku_{xt}(x,t) + \mathcal{O}(k^2) + \frac{1}{2}hu_{xx}(x,t) + \mathcal{O}(hk) + \mathcal{O}(h^2)$$
$$= u_x(x,t) + \alpha ku_{xx}(x,t) + \mathcal{O}(k^2) + \frac{1}{2}hu_{xx}(x,t) + \mathcal{O}(hk) + \mathcal{O}(h^2)$$

于是，我们有

$$A + B - \alpha(C+D) = \mathcal{O}(k^2) + \mathcal{O}(hk) + \mathcal{O}(h^2)$$
$$= \mathcal{O}(h^2 + k^2)$$

■

9.7.4 伽辽金法

伽辽金法也可应用于双曲型问题．用伴随初始条件和边界条件的单个一阶微分方程：

664

$$\begin{cases} u_t = \alpha u_x \\ u(x,0) = g(x) & (0 \leqslant x \leqslant 1) \\ u(0,t) = u(1,t) = 0 & (t > 0) \end{cases} \tag{15}$$

来说明具体方法．首先选择某些 x 的基函数，例如 w_1，w_2，…，w_n．我们尝试用下列形式的函数

$$u(x,t) = \sum_{j=1}^{n} v_j(t)w_j(x) \tag{16}$$

求解我们的问题．可用函数 v_1，v_2，…，v_n 来协助这项工作．把试验函数(16)代入到微分方程中，得到

$$\sum_{j=1}^{n}[v'_j(t)w_j(x) - \alpha v_j(t)w'_j(x)] = 0 \tag{17}$$

照例在接下来的这个方法中，我们不期望方程(17)是相容的．即我们不期望找到函数 v_j 使这个方程成立．当然，其理由是边值问题(15)的解可能不能像(16)式那样根据选择的函数 w_j 来表达．在伽辽金法(以及其他类似的方法)中，寻求的是方程(17)的近似解．同时，必须考虑到

(15)式中的边界和初始条件. 为了简单起见，在 0 及 1 上的每个基函数 w_j 为 0. 由试验函数(16)知，齐次边界条件将自动满足. 下面，对(17)式两边用 w_i，$1\leqslant i\leqslant n$ 作内积，得到

$$\sum_{j=1}^{n}[v_j'(t)\langle w_j,w_i\rangle-\alpha v_j(t)\langle w_j',w_i\rangle]=0\qquad(1\leqslant i\leqslant n)\tag{18}$$

这里记号 $\langle f,g\rangle=\int_0^1 f(x)g(x)\mathrm{d}x$. (18) 式是 n 个未知函数 v_j 的 n 个线性齐次微分方程的方程组. 我们注意到选择 $\{w_1,w_2,\cdots,w_n\}$ 作为区间[0,1]上的**标准正交系**有很大的优点. 假定(18) 式采取形式

$$V'=AV\tag{19}$$

其中 $V=(v_1,\ v_2,\ \cdots,\ v_n)^{\mathrm{T}}$ 而 A 是 $n\times n$ 矩阵，其元素是

$$a_{ij}=\alpha\langle w_j',w_i\rangle$$

(15)式中的初始条件现在变成

$$\sum_{j=1}^{n}v_j(0)w_j(x)=g(x)\tag{20}$$

另外，此式也许不可能精确地满足，并且如果是这样，则我们可选择 $v_j(0)$ 使得(20)式在 L^2 范数意义下尽可能地接近满足. 因为函数 w_j 已假定构成标准正交系，所以这意味着

665

$$v_j(0)=\langle g,w_j\rangle\tag{21}$$

(20)式提供求解方程组(19)的初始条件. 正如我们从 8.11 节知道，这个方程组的解是

$$V(t)=\mathrm{e}^{tA}V(0)$$

一个方便的标准正交系由在 0 和 1 上为零的函数

$$w_j(x)=2^{-1/2}\sin\pi jx$$

组成. 另一个标准正交系可对函数序列 $x\mapsto(x^2-x)x^j$，$j=1,\ 2,\ \cdots,\ n$ 用格拉姆-施密特过程来构造.

习题 9.7

1. 证明：若 $u_t=\alpha u_x$，则对 $i=0,\ 1,\ 2,\ \cdots$，

$$\frac{\partial^n u}{\partial t^n}=\alpha^n\frac{\partial^n u}{\partial x^n}$$

并且证明向量的情况.

2. 根据(3)式导出一个三阶逼近方法.
3. 提供导致(6)式和(7)式以及稳定性条件 $k|\alpha|\leqslant h$ 的论证中所有的细节.
4. 对方程(1)的数值过程

$$\frac{1}{2k}(v_{j,n+1}-v_{j,n-1})=\frac{\alpha}{2h}(v_{j+1,n}-v_{j-1,n})$$

进行稳定性分析.

5. 研究方程(1)的欧拉方法的稳定性：

$$v_{j,n+1}=v_{j,n}+\frac{\alpha k}{2h}(v_{j+1,n}-v_{j-1,n})$$

6. 设 u 和 v 是两个复数，$u=x+\mathrm{i}y$，$v=a+\mathrm{i}b$. 证明下列不等式等价：

a. $|u+v|\leqslant|u-v|$

b. $xa+yb\leqslant 0$

由此证明(14)式.

7. 参考温德罗夫方法并说明截断误差中主项是

$$\frac{1}{12}(ah^2-a^3k^2)u_{xxx}(x,t)$$

计算机习题 9.7

1. 编写(4)式的拉克斯-温德罗夫方法的程序. 假定数值解是在由 $a\leqslant x\leqslant b$ 和 $t=T$ 所定义的线段上计算的，其中 a，b 和 T 是给定的. 当然，f 和 α 也是给定的. 用户希望指定步长.
2. (续)对问题 $u_t=2u_x$，$u(x,0)=(1-x)^2$，测试上题中编写的程序. 取 $h=0.02$，$k=0.01$，$a=1$，$b=2$ 和 $T=1$. 比较数值解与真解.

666

9.8 多重网格方法

微分方程数值解的**多重网格方法**是基于离散化和接着用有限差分公式逼近导数. 这个方法的卓越性质是在区域上使用一些从粗到细排列的不同网格. 在粗网格上的数值解可很快地计算，但它将具有低精度. 虽然这样，但它作为细网格上迭代解的初始点可能是有用的. 这仅仅是多重网格方法的一个方面，我们从一个简单的例子入手说明它如何工作.

9.8.1 作为说明的例子

考察下列两点边值问题

$$\begin{cases} u''(x)=f(x) \\ u(0)=u(1)=0 \end{cases} \tag{1}$$

此例仅用作说明：多重网格方法的实力仅当我们开始将它用于偏微分方程时才呈现出来.

若选定步长 h，则可用标准的方式离散化问题(1)：

$$h^{-2}(v_{j-1}-2v_j+v_{j+1})=f_j \qquad (1\leqslant j\leqslant n) \tag{2}$$

这里我们采用下列这些定义：

$$h=\frac{1}{n+1} \qquad x_j=jh \qquad v_j\approx u(x_j) \qquad f_j=f(x_j) \qquad v_0=v_{n+1}=0$$

不难直接求解 $n\times n$ 的方程组(2)，但是考虑到求解偏微分方程的必要性，我们选择用迭代法处理(2)，为此选择高斯-赛德尔方法. 当然，如果提供一个好的初始向量，迭代法将很快产生一个满意的解. 对即将处理的问题得到初始点的一个方法是在粗网格上求解同样的问题，因而，可设想方程组(2)用较大的步长 h 写出. 这个方程组将有较少的方程，但方程有同样的结构. 方程组具有形式

$$\begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} -h^2f_1 \\ -h^2f_2 \\ -h^2f_3 \\ \vdots \\ -h^2f_n \end{bmatrix}$$

因此，这个方程组也可用高斯-赛德尔法迭代求解. 若我们选择这样做，则一定希望从一个好

的初始向量开始，而这样的初始点仍可利用粗网格得到，等等．因此，合乎逻辑的是开始用粗网格．此时方程组(2)处于最简单的情况，即 $n=1$．当 $n=1$，$h=1/2$ 时，方程组中只有一个方程．它的解是

667

$$v_0 = 0 \qquad v_1 = -\frac{1}{8}f(\frac{1}{2}) \qquad v_2 = 0$$

现在我们准备转换到细网格点．用 2 除 h，用 $2n+1$ 代替 n，建立新的方程组．因此，$n=3$，$h=1/4$．我们临时使用 w 作为这个新方程组中包含的向量．用 v 赋适当的初始值于分量 w_0，w_1，w_2，w_3 和 w_4．图 9-13 中所示为 v 和 w 之间关系的略图．根据向量 v 携带的信息赋值给向量 w 可用许多方式实现．通常使用的是一个简单的插值格式．这里作为说明，我们将在适当的分量上复制 v 的值到 w 上，其他的分量利用平均值．因此，使用的 5 个式子是：

w_0	w_1	w_2	w_3	w_4
•	•	•	•	•
v_0		v_1		v_2
•		•		•
0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1

图 9-13　说明多重网格从粗网格到细网格的例子

$$w_0 = v_0 \qquad w_2 = v_1 \qquad w_4 = v_2 \qquad w_1 = (v_0 + v_1)/2 \qquad w_3 = (v_1 + v_2)/2$$

这样，就抛弃了老的 v 向量并用 w 代替它．下面我们用刚才构造的初始向量 v 来执行少许高斯-赛德尔方法迭代．然后通过将 h 减半从而改变 n 来形成新的网格点．以后的过程重复进行．

```
input m, k
h←1/2; n←1
v0←0; v2←0; v1←-(1/8)f(h)
for i=2 to m do
    h←h/2; n←2n+1
    for j=0 to (n+1)/2 do
        w2j←vj
    end do
    for j=1 to (n+1)/2 do
        w2j-1←(vj-1+vj)/2
    end do
    for j=0 to n+1 do
        vj←wj
    end do
    for p= 1 to k do
        for j=1 to n do
        vj←[vj-1+vj+1-h^2 f(jh)]/2
    end do
    end do
    output (vi)
end do
```

668

在伪代码中，f 是出现在原问题(1)中的函数．使用的网格点数(包括第 1 个)是 m，k 是在每个网格上执行高斯-赛德尔迭代的次数．从 v 到 w 的信息转移称为**插值阶段**．我们用公式

$$w_{2j} = v_j \qquad w_{2j-1} = (v_{j-1} + v_j)/2 \tag{3}$$

表示算法的一般步．在伪代码中，起始于“**for** $p=1$ **to** k **do**”的循环是执行高斯-赛德尔迭代 k

步. 因此，方程(2)是求解第 j 个未知量，导致

$$v_j=(v_{j-1}+v_{j+1}-h^2 f_j)/2 \tag{4}$$

这个公式用于更新每个 v_j，整个迭代过程执行 k 次.

从 v 产生 w，然后用 w 代替 v 的插值过程可如习题 9.8.1 中暗示的那样更为简单地获得. 代码仅仅用作教学目的，但它可用于实验——特别是其解是已知的问题，因为可以考察实际的解与数值解之间的差别. 在 $f(x)=\cos x$ 的情况下这是可行的，相关的两点边值问题的解是

$$u(x)=-\cos x+x(\cos 1-1)+1 \tag{5}$$

设 m(格点改进数)是 6，并设 k (高斯-赛德尔迭代数)是 3，我们发现计算解和真解之间相差大约 10^{-3}. 在最后的格点上 h 的值是 1/64，且 $h^2\approx 2\times 10^{-4}$.

9.8.2 误差的阻尼

多重网格方法另一个要素是系统地从细网格前进到较粗网格. 其方向与我们前面说明的相反，在该方向上前面粗网格的信息向上转移从而为较细网格上的迭代提供初始点. 在计算过程的后面的点上结合较粗网格的原因是在对应于粗网格的迭代中往往有效地阻尼低频误差. 相反，在细网格上的迭代中往往有效地阻尼高频误差. 多重网格方法便利用了这个重要特性，并且在很大程度上，这是方法成功的原因.

利用微分方程(1)的一个简单的数值实验指出不同频率误差的阻尼. 考虑齐次方程

$$\begin{cases} u''=0 \\ u(0)=u(1)=0 \end{cases} \tag{6}$$

669

显然它的解是 $u(x)\equiv 0$. 我们用(4)式描述的高斯-赛德尔迭代求解这个问题，在方程中取 $f=0$ 并用正弦曲线函数

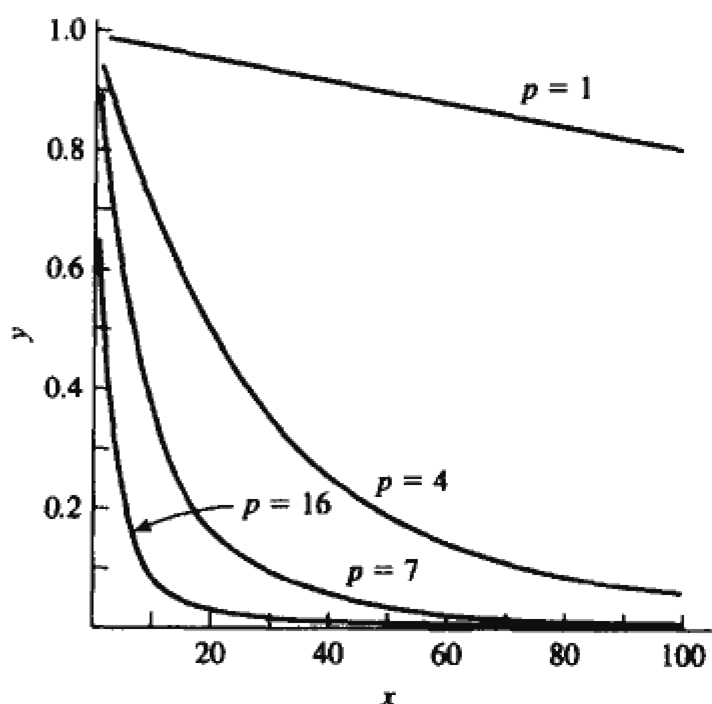
$$v_j=\sin\left(\frac{jp\pi}{n+1}\right)\qquad (0\leqslant j\leqslant n+1)$$

作为起始点. 这里 n(如前面所用的)是区间(0，1)上的内结点数. 它控制网格的精细度. 在这个实验中，因为问题的解是 $u\equiv 0$，所以向量 v 被认为是误差.

参数 p 控制初始误差的频率. 给出利用 4 个给定的频率执行 k 次迭代的伪代码. 在每次高斯-赛德尔迭代结束时计算向量的范数.

```
input n, k, p1, p2, p3, p4
for p=p1, p2, p3, p4 do
    for j=0 to n+1 do
        vj=sin((jpπ)/(n+1))
    end do
    for i=1 to k do
        for j=1 to n do
          vj=(vj-1+vj+1)/2
    end do
    ρi=‖v‖∞=max1≤j≤n |vj|
    output p, i, ρi
    end do
end do
```

当这个程序用 $n=63$，$k=100$ 和 $p=1$，4，7，16 运行时，得到如图 9-14 中所示的结果. 显然高频的误差通过迭代被迅速地阻尼，而低频误差仅仅被轻微地阻尼.



670

图 9-14 在 100 次迭代中误差的缩减

9.8.3 分析

为分析迭代对误差的阻尼作用，我们选择雅可比方法作为说明. 当雅可比迭代用于方程组 $Ax=b$ 时，迭代公式是

$$x^{(k+1)}=(I-D^{-1}A)x^{(k)}+D^{-1}b \tag{7}$$

其中 D 是 A 的对角元构成的矩阵. 在本节的例子中，(2)式涉及的矩阵 A 是

$$A=\begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 2 \end{bmatrix}$$

而右边是 $b=h^2 f_j$. 所以迭代矩阵是

$$G\equiv I-D^{-1}A=\begin{bmatrix} 0 & \frac{1}{2} & 0 & \cdots & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & \cdots & 0 \\ 0 & \frac{1}{2} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

显然

$$G=I-\frac{1}{2}A \tag{8}$$

由 9.1 节中的引理 1，A 的特征值是

$$\mu_j = 2 - 2\cos\frac{j\pi}{n+1} \qquad (1 \leqslant j \leqslant n) \tag{9}$$

对应于 μ_j的特征向量是

$$v^{(j)} = \left(\sin\frac{j\pi}{n+1}, \sin\frac{2j\pi}{n+1}, \ldots, \sin\frac{nj\pi}{n+1}\right) \tag{10}$$

由(8)式，迭代阵 G 的特征值λ_j 是

$$\lambda_j = 1 - \frac{1}{2}\mu_j = \cos\frac{j\pi}{n+1} \qquad (1 \leqslant j \leqslant n)$$

所以迭代阵之谱半径是

$$\rho(G) = \cos\frac{\pi}{n+1} \approx 1 - \frac{1}{2}\left(\frac{\pi}{n+1}\right)^2 = 1 - \frac{\pi^2}{2}h^2$$

671

现在，我们开始引出结论. 首先，因为 $\rho(G)<1$，所以由 4.6 节定理 5 知，雅可比迭代收敛. 其次，我们注意到对小的 h 值，谱半径接近于 1，这说明较精细的网格收敛非常慢. 第三，我们可利用(10)式中 A 的特征值去理解误差的阻尼. 向量 $v^{(j)}$也是 G 的特征向量，并且它们构成 $\mathbb{R}^n$ 的一个基. 因此，任何在数值解中出现的误差必定是 $v^{(1)}$，$v^{(2)}$，…，$v^{(n)}$的线性组合. 考虑误差的单个分量 $v^{(j)}$就足够了. 显然，用 G 对 $v^{(j)}$迭代 k 次将产生$\lambda_j^k v^{(j)}$，因为 $|\lambda_j|<1$，所以当 $k\to\infty$时，它将收敛于 0. 但是这个**阻尼作用**的强度当 $|\lambda_j|$ 最小时将较大. $|\lambda_j|$ 小的值对应于范围 $1\leqslant j\leqslant n$ 中间的 j，然而对 $j\approx 1$ 或 $j\approx n$，我们有 $|\lambda_j|\approx 1$.

9.8.4 限制和网格校正

整个多重网格方法包含许多步，每一步中在一个网格上得到的结果被传送到下一组计算的另一个网格. 我们已经指出如何从粗网格向上计算到细网格. 在进行向下计算时，假定利用方程组(2)的一个近似解向量 v^i. 不用当前的 h 值在方程组(2)中迭代，而是决定下降到粗网格做迭代，因为在粗网格上将处理得更快. 我们的目标是对 v^i增加适当的校正来改善 v^i. 若 e^i是校正项而方程组(2)中的矩阵用 A^i 表示，则需要解方程

$$A^i(v^i + e^i) = f^i \tag{11}$$

这里 f^i 表示当前网格上的函数 f. 记原来的方程为

$$A^i e^i = f^i - A^i v^i = r^i$$

其中 r^i 是由近似解 v^i造成的残差. 通过传送到较粗网格并用少量高斯-赛德尔方法的迭代求解方程组 $A^i e^i = f^i$. 在粗网格上，有方程组

$$A^{i-1}e^{i-1} = r^{i-1} \tag{12}$$

用细网格到粗网格上信息的某个初等变换从 r^i 得到向量 r^{i-1}. 这样做的一个方法是记

$$r_j^{i-1} = r_{2j}^i \tag{13}$$

一个更精细的方法是利用 r^i 中的全部信息通过加权平均给出

$$r_j^{i-1} = \frac{1}{4}r_{2j-1}^i + \frac{1}{2}r_{2j}^i + \frac{1}{4}r_{2j+1}^i \tag{14}$$

不管使用什么公式，这个过程称为**限制**.

672

对(12)式作少量高斯-赛德尔方法的迭代以后，我们将有向量 e^{i-1}. 利用前面讨论过的插值过程，我们得到向量 e^i，并用运算

$$v^i \leftarrow v^i + e^i$$

更新 v^i. 这个过程称为**粗网格校正格式**.

9.8.5 V 循环算法

我们最终的目标是描述多重网格算法中所谓的 **V 循环**. V 循环的名称可从图 9-15 中的简图中得到，此图描述了计算过程. 图中的每个圆表示在一个网格上执行的一块计算. 从最细的网格上开始计算. 经过少量迭代后，算出残差 r，并将控制权转移到一个较粗的网格上，在粗网格上处理形式为 $Az=r$ 的方程. 这个过程是重复的，穿过一连串的网格向下传递直到得到最粗网格上的一个方程组. 这个方程组通常是精确求解的，因为它只由少数几个方程，可能刚好是一个方程所组成. 在图的向上部分中，通过插值过程以及执行附加的迭代由较粗网格传递信息到较细网格上.

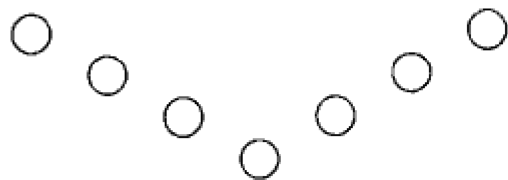


图 9-15 说明多重网格中 V 循环的例子

为给出 V 循环的正规描述，首先给网格编号：$\mathcal{G}^1$，$\mathcal{G}^2$，…，$\mathcal{G}^m$，其中 $\mathcal{G}^1$ 表示最粗网格而 $\mathcal{G}^m$ 表示最细网格. 在网格 $\mathcal{G}^i$ 上，矩阵 A^i 是指定的. 我们希望求解的问题是 $A^m v^m = f^m$；这是给定的边值问题在最细网格上的离散形式. 因而 f^m 是输入向量. 然而，对 $i=m-1$，$m-2$，…，1，f^i 是逐次计算的. 于是，V 循环的步骤如下：

1. 置 i 等于 m，并根据边值问题把数据放入 f^m 中，把最好的可利用的猜测放入 v^m 中.

2. 对方程组 $A^i v^i = f^i$ 应用 k 次迭代法，用当前的 v^i 作为初始值. 计算残差 $r^i \leftarrow f^i - A^i v^i$. 应用限制算子 R_i：$f^{i-1} \leftarrow R_i r^i$. 从 i 中减去 1.

3. 若 $i=1$，转到第 4 步. 若 $i>1$，转到第 2 步.

4. 精确地求解 $A^1 v^1 = f^1$.

5. i 加 1. 应用扩张算子 E_i 并再加上校正项 $v^i \leftarrow v^i + E_i v^{i-1}$. 对方程组 $A^i v^i = f^i$ 应用 k 次迭代法，用当前的 v^i 作为初始值.

6. 若 $i=m$，输出 v^m 并停止. 若 $i<m$，转到第 5 步.

673

下面给出我们的模型问题在多重网格方法中应用 V 循环的伪代码.（为节省空间，对循环使用缩写记号.）在网格 $\mathcal{G}^i$ 中，步长是 $h=2^{-i}$. 临时工作区域 w 用于存放中间量值. 在代码中使用的限制算子是最简单的一种，如(13)式中那样，由此把偶-指标分量直接复制到下一个粗网格的向量中去. 代码中其他一切均与前面列举的算法一致. 使用高斯-赛德尔迭代法. 在这个算法中，向量 v_j^i 的元素可存放在一个二维数组中，例如，$V(J, I) \leftarrow v_j^i$，使得数组的列对应不同层次上的网格值.

伪代码用于前面提到的模型问题

$$\begin{cases} u'' = \cos x \\ u(0) = u(1) = 0 \end{cases}$$

的各种数值实验. (5)式是这个问题的解. 例如，我们置 $m=7$，其在最细网格上对应于 $h=$

1/128. 然后，用各种 k 的值确定这个参数对误差的影响. 这里误差与 $\max_{0\leqslant j\leqslant n+1}|u(x_j)-v_j^m|$ 有关，其中 u 是真解，$x_j=jh$，v^m 是算法用 $h=2^{-m}$ 在最细网格上产生的近似解. 每次高斯-赛德尔迭代数增加 1 次，新的误差大约是前面误差的 2/5. 对 $3\leqslant k\leqslant 8$，这个结果成立. 进一步增加 k 致使误差不太显著地减缩，对 $9\leqslant k\leqslant 12$，减缩因子大约是 0.6，0.7，0.8 和 0.9.

```
input m, k
n ← 2^m − 1
h ← 1/(n+1)
v_j^i ← 0;  f_j^i ← 0     (1≤i≤m, 0≤j≤n+1)
f_j^m ← f(jh)     (1≤j≤n)
for i = m to 2 step −1 do
    for p = 1 to k do
        v_j^i ← [v_{j−1}^i + v_{j+1}^i − h^2 f_j^i]/2     (1≤j≤n)
    end do
    w_j ← f_j^i − [v_{j−1}^i − 2v_j^i + v_{j+1}^i]/h^2     (1≤j≤n)
    f_j^{i−1} ← w_{2j}     (1≤j≤(n−1)/2)
    h ← 2h
    n ← (n−1)/2
end do
v_1^1 = −f(1/2)/8
for i = 2 to m do
    h ← h/2
    n ← 2n+1
    w_{2j} ← v_j^{i−1}     (0≤j≤(n+1)/2)
    w_{2j−1} ← [v_{j−1}^{i−1} + v_j^{i−1}]/2     (1≤j≤(n+1)/2)
    v_j^i ← v_j^i + w_j     (0≤j≤n+1)
    for p = 1 to k do
        v_j^i ← [v_{j−1}^i + v_{j+1}^i − h^2 f_j^i]/2     (1≤j≤n)
    end do
end do
output v_j^m     (0≤j≤n+1)
```

674

9.8.6 运算量

为估计多重网格算法的计算成本，我们考察代码中给出的 V 循环和 V 循环中算法的运算量.

在过程的向下部分，使用 m 个不同的网格. 在网格 $\mathcal{G}^i$ 上，有 2^i 个点，2^i 个未知量和 2^i 个方程. 一个变量的每次更新需要高斯-赛德尔方法中的 4 次运算. 因此，在每个网格上，迭代中需要进行 $4k2^i$ 次运算. 残差计算增加 $5\cdot 2^i$ 次运算而限制算子不增加什么. 因此，在第 i 个网格上，使用 $(4k+5)2^i$ 次运算. 所有 m 个网格的总运算次数是

$$\sum_{i=1}^{m}(4k+5)2^i \approx (4k+5)2^{m+1}=(8k+10)2^m$$

类似地计算 V 循环的向上部分，得到总计大约为 $(8k+4)2^m$ 次计算量. 于是对整个 V 循环包含大约 $16(k+1)2^m$ 次运算.

对于在单位正方形上像$\nabla^2 u=f(x, y)$那样的二维问题，相应的次数是多少？在常见的离散化中每个方程包含 5 个未知量，用高斯-赛德尔方法更新每个变量大约需要 6 次运算. 因为有两个变量，所以在第 i 个网格上的变量数为$(2^i)^2$. 因此，我们看到对计算的主要影响是把 2^m 变到 4^m. 乘 4^m 的因子比 $16(k+1)$大，但它仍然是 k 的线性函数.

前面的注记说明计算量是 m(网格数)的指数函数和 k(指定的迭代数)的线性函数. 这个结论对任意维数均成立.

习题 9.8

插值阶段从 v 计算 w，然后由 w 代替 v. 求仅仅利用 v 数组做此项工作的一个有效的代码.

计算机习题 9.8

1. 重复课本中介绍的设计指出缩小不同频率误差的数值试验.
2. 重复课本中在连续的细网格上求解模型问题的数值试验.
3. 重复课本中利用 V 循环算法的数值试验.
4. 对下列二维问题：

$$\begin{cases} u_{xx}+u_{yy}=f(x,y) & (0<x<1, 0<y<1) \\ u(x,y)=0 & \text{在边界上} \end{cases}$$

[675]

编写 V 循环算法程序. 当 $f(x, y)=2x(x-1)+2y(y-1)$时，测试你的程序. 真解是 $u(x, y)=xy(1-x)(1-y)$.

5. 在本节第 1 个代码中，调整循环使得 v_0 和 v_{n+1}保持常数，而不是来回复制.
6. 利用本节第 1 个代码，执行一些数值试验来确定是否存在高斯-赛德尔迭代次数 k 的最优值.
7. 推广本节中的代码使之适合问题 $u''(x)=f(x)$，$u(a)=\alpha$，$u(b)=\beta$.

9.9 泊松方程的快速方法

两个变量的泊松方程是

$$u_{xx}+u_{yy}=f(x,y) \tag{1}$$

在涉及此方程的典型的物理问题中，我们寻找在某个指定的开区间 Ω 中满足方程(1)，并且在 Ω 的边界(由 $\partial\Omega$ 表示)上满足指定条件的函数 u. 函数 f 定义在 Ω 上.

近年来，傅里叶分析已被应用于得到求解此类边值问题的快速算法. 这些新方法利用快速傅里叶变换. 下面用一个简单的模型问题来说明导出这些新算法的新方法.

9.9.1 模型问题

模型问题是：

$$\begin{cases} \Omega=\{(x,y): 0<x<1, 0<y<1\} \\ u_{xx}+u_{yy}=f(x,y) & \text{在 } \Omega \text{ 内} \\ u(x,y)=0 & \text{在 } \partial\Omega \text{ 上} \end{cases} \tag{2}$$

我们着手离散化，取

$$h=\frac{1}{n+1} \qquad x_i=ih \qquad y_j=jh \qquad (0\leqslant i,j\leqslant n+1)$$

对问题(2)，引入一个常见的近似

$$v_{ij} \approx u(x_i, y_j) \qquad f_{ij} = f(x_i, y_j)$$

问题的离散化形式是

$$h^{-2}(v_{i+1,j} - 2v_{ij} + v_{i-1,j}) + h^{-2}(v_{i,j+1} - 2v_{ij} + v_{i,j-1}) = f_{ij} \tag{3}$$

在(3)式中，i 和 j 的范围是$\{1, 2, \cdots, n\}$. 在(3)式中大多数的项是未知的，但是因为边界条件，我们要求

$$v_{0j} = v_{n+1,j} = v_{i0} = v_{i,n+1} = 0 \tag{4}$$

676

此时传统的做法是用迭代法解方程组(3). 这里有 n^2 个方程和 n^2 未知量. 用逐次超松弛求解此方程组的计算工作量是 $\mathcal{O}(n^3\log n)$. 另一个含有快速变换的方法可使这个工作量降到 $\mathcal{O}(n^2\log n)$.

9.9.2 快速傅里叶正弦变换

寻找下列形式的方程组(3)的解：

$$v_{ij} = \sum_{k=1}^{n} \hat{v}_{kj} \sin ik\phi \qquad (0 \leqslant i, j \leqslant n+1) \tag{5}$$

其中 $\phi = \pi/(n+1)$. 这里数 $\hat{v}_{kj}$ 是我们希望确定的未知数. 它们表示函数 v 的傅里叶正弦变换. 一旦 $\hat{v}_{kj}$ 被确定，就可用快速傅里叶正弦变换有效地计算 v_{ij}.

若由(5)式把 v_{ij} 代入到(3)式中，结果是

$$\begin{aligned}&\sum_{k=1}^{n} \hat{v}_{kj}[\sin(i+1)k\phi - 2\sin ik\phi + \sin(i-1)k\phi] \\ &+ \sum_{k=1}^{n} \sin ik\phi[\hat{v}_{k,j+1} - 2\hat{v}_{kj} + \hat{v}_{k,j-1}] = h^2 f_{ij}\end{aligned} \tag{6}$$

现在，利用引理 2 中的三角恒等式简化第 1 个和式. 同时，引入 f_{ij} 的正弦变换：

$$f_{ij} = \sum_{k=1}^{n} \hat{f}_{kj} \sin ik\phi$$

结果是

$$\begin{aligned}&\sum_{k=1}^{n} \hat{v}_{kj}(-4\sin ik\phi)\left(\sin^2 \frac{k}{2}\phi\right) \\ &+ \sum_{k=1}^{n} \sin ik\phi(\hat{v}_{k,j+1} - 2\hat{v}_{kj} + \hat{v}_{k,j-1}) = h^2 \sum_{k=1}^{n} \hat{f}_{kj} \sin ik\phi\end{aligned} \tag{7}$$

由引理 1 知，具有元素 $\sin ij\phi$ 的矩阵非奇异. 所以，可以从(7)式导出

$$\hat{v}_{kj}\left(-4\sin^2 \frac{k\phi}{2}\right) + \hat{v}_{k,j+1} - 2\hat{v}_{kj} + \hat{v}_{k,j-1} = h^2 \hat{f}_{kj} \tag{8}$$

初看起来(8)式好像是另一个有 n^2 未知量和 n^2 个方程的方程组，它仅仅稍稍不同于原来的方程组(3). 但仔细的检查显示：在(8)式中，k 可取固定的，并且因为所得的 n 个方程的方程组是三对角的，所以它容易直接求解. 因此，对固定的 k，(8)式中的未知量构成$\mathbb{R}^n$ 中的一个向量

677

$$(\hat{v}_{k1}, \hat{v}_{k2}, \ldots, \hat{v}_{kn})$$

上面所用的方法把原来的 n^2 个方程的方程组**拆开**成 n 个方程组，而每个方程组有 n 个方程. n 个方程的三对角方程组可用 $\mathcal{O}(n)$次运算求解(事实上，需要少于 $10n$ 次运算). 因此，我们可

用 $10n^2$ 次运算求解 n 个三对角方程组. 快速傅里叶正弦变换对 n 个分量的向量使用 $\mathcal{O}(n\log n)$ 次运算. 因此, 在快速泊松方法中总的计算量是 $\mathcal{O}(n^2\log n)$.

9.9.3 附加的细节

有必要提出前面讨论中的一些细节. 首先, 从(5)式中观察边界条件

$$v_{0j} = v_{n+1,j} = 0 \qquad (0 \leqslant j \leqslant n+1)$$

会自动地满足而不对系数 $\hat{v}_{kj}$ 作任何限制. 然而, 其余的边界条件

$$v_{i0} = v_{i,n+1} = 0$$

将不满足, 除非我们补充两个等式:

$$\sum_{k=1}^{n} \hat{v}_{k0} \sin ik\phi = \sum_{k=1}^{n} \hat{v}_{k,n+1} \sin ik\phi = 0 \qquad (0 \leqslant i \leqslant n+1) \tag{9}$$

(9)式表明两个向量

$$(\hat{v}_{10}, \hat{v}_{20}, \ldots, \hat{v}_{n,0}) \text{ 和 } (\hat{v}_{1,n+1}, \hat{v}_{2,n+1}, \ldots, \hat{v}_{n,n+1})$$

必须正交于(引理 1)那个非奇异矩阵的行. 因此, 必须定义

$$\hat{v}_{k0} = \hat{v}_{k,n+1} = 0 \qquad (1 \leqslant k \leqslant n)$$

引理 1(对称正交矩阵引理) 元素为

678

$$a_{kj} = (2/n)^{1/2} \sin \frac{kj\pi}{n} \qquad (1 \leqslant k \leqslant n-1, 1 \leqslant j \leqslant n-1)$$

的 $(n-1)\times(n-1)$ 矩阵 A 是对称和正交的. 因此, $A^2 = I$.

证明 我们计算 A^2 的一般元素:

$$\begin{aligned} A_{kj}^2 &= \sum_{v=1}^{n-1} a_{kv} a_{vj} = \frac{2}{n} \sum_{v=1}^{n-1} \sin \frac{kv\pi}{n} \sin \frac{jv\pi}{n} \\ &= \frac{1}{n} \sum_{v=1}^{n-1} \left[\cos \frac{v(k-j)\pi}{n} - \cos \frac{v(k+j)\pi}{n} \right] \\ &= \frac{1}{n} \mathrm{Re}\left[\sum_{v=0}^{n-1} (e^{iv\phi} - e^{iv\theta}) \right] \end{aligned} \tag{10}$$

其中 $\phi = (k-j)\pi/n$, $\theta = (k+j)\pi/n$.

若 $k=j$, 则 $\phi=0$ 且 $\theta = 2k\pi/n$, 因为 $1 \leqslant k \leqslant n-1$, 所以 θ 不是 2π 的倍数. 因此,

$$A_{kk}^2 = \frac{1}{n} \mathrm{Re}\left[n - \frac{e^{in\theta} - 1}{e^{i\theta} - 1} \right] = 1$$

这里我们注意到 $e^{in\theta} = e^{i2k\pi} = 1$.

若 $k \neq j$, 则 ϕ 和 θ 都不是 2π 的倍数. (10)中的几何级数可利用通常的公式求和. 结果是

$$A_{kj}^2 = \frac{1}{n} \mathrm{Re}\left[\frac{e^{in\phi} - 1}{e^{i\phi} - 1} - \frac{e^{in\theta} - 1}{e^{i\theta} - 1} \right]$$

若 $k-j$ 是偶数, 则 $k+j$ 也是偶数, 并且此时 $e^{in\phi} = e^{in\theta} = 1$. 因此, $A_{kj}^2 = 0$. 另一方面, 若 $k-j$ 是奇数, 则 $k+j$ 也是奇数. 此时 $e^{in\phi} = e^{in\theta} = -1$. 因此, 我们要证明

$$\mathrm{Re}\left[\frac{-2}{e^{i\phi} - 1} - \frac{-2}{e^{i\theta} - 1} \right] = 0 \tag{11}$$

为了证明这个结论，首先注意到对 $z\neq0$

$$\mathrm{Re}\left[\frac{1}{z}\right]=\mathrm{Re}\left[\frac{\bar{z}}{z\bar{z}}\right]=\frac{\mathrm{Re}[z]}{|z|^2}$$

然后，应用这个关系于 $z=\mathrm{e}^{i\phi}-1$，得到

$$\mathrm{Re}\left[\frac{1}{\mathrm{e}^{i\phi}-1}\right]=\frac{\cos\phi-1}{(\cos\phi-1)^2+\sin^2\phi}=\frac{\cos\phi-1}{2-2\cos\phi}=-\frac{1}{2}$$

当然，当 $\phi=\theta$ 时，上式也成立．因此，(11)式左边的两项是相同的． ■

引理 2（$\sin(A+B)$的引理）

$$\sin(A+B)-2\sin A+\sin(A-B)=-4\sin A\sin^2\frac{B}{2}$$

679

证明 利用下列熟悉的恒等式

$$\sin(A\pm B)=\sin A\cos B\pm\cos A\sin B$$

$$1-\cos 2A=2\sin^2 A$$

■

计算机习题 9.9

用各种 n 的值以及你的程序库中执行快速泊松求解过程的计算机程序来解模型问题(2)．

680

第 10 章 线性规划及其相关论题

10.1 凸性和线性不等式

线性不等式的理论非常类似于更为常见的线性方程的理论．这里我们将讨论这个主题的基础部分并指出它的某些应用．

10.1.1 基本概念

在这个主题中所使用的一切向量和矩阵都是实的（非复的），因为理论本质上利用实线的有序结构．对$\mathbb{R}^n$ 中的两个点（向量）x 和 y，我们记

$$x \geqslant y \quad \text{当且仅当} \quad x_i \geqslant y_i (1 \leqslant i \leqslant n)$$

类似地，我们用分量不等式来定义 $x \leqslant y$，$x > y$ 或 $x < y$．特别应该注意 $x > y$ 与 $x \geqslant y$ 且 $x \neq y$ 是不同的（原因是 $x \neq y$ 并不意味着对一切 i，$x_i \neq y_i$，而是对某个 i，$x_i \neq y_i$）．

n 个变量的 m 个弱线性不等式系统可以写成

$$Ax \geqslant b \tag{1}$$

其中 A 是 $m \times n$ 矩阵，x 是 n 维向量，b 是 m 维向量，在这个系统中单个的不等式是

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n \geqslant b_i \qquad (0 \leqslant i \leqslant m)$$

或

$$A_{\text{row}_i} x \geqslant b_i \qquad (1 \leqslant i \leqslant n)$$

其 A_{row_i} 是矩阵 A 的第 i 个行向量．这样的不等式系统涉及的一个基本问题是它是否**相容**．换句话说，是否存在 x 使得 $Ax \geqslant b$? 如果它是相容的，则我们想得到求解的算法．更一般的线性不等式系统可能是

681

$$\begin{cases} \sum_{j=1}^{n} a_{ij}x_j \geqslant b_i & (1 \leqslant i \leqslant m_1) \\ \sum_{j=1}^{n} a_{ij}x_j > b_i & (m_1 + 1 \leqslant i \leqslant m) \end{cases}$$

但为了简单起见，我们考虑系统(1)．

为了看看希望得到什么，我们考虑一个小的例子，其中 $n=2$，$m=3$：

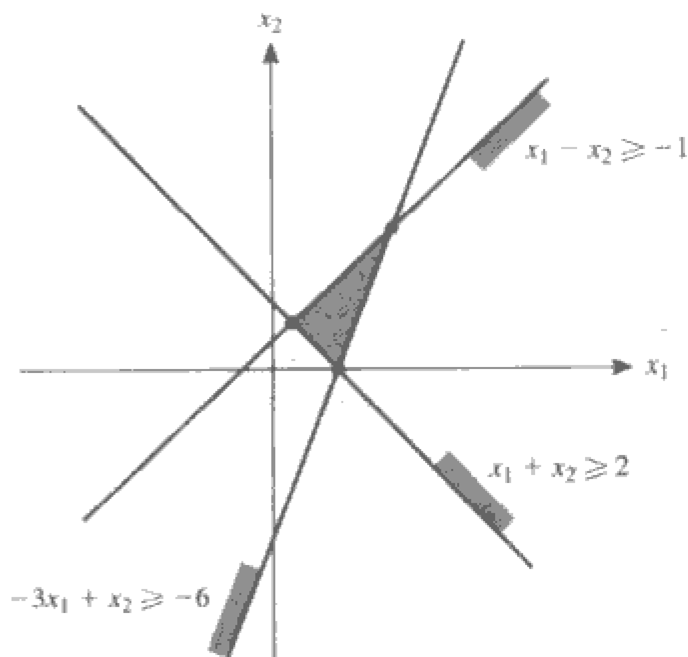
$$\begin{cases} x_1 + x_2 \geqslant 2 \\ x_1 - x_2 \geqslant -1 \\ -3x_1 + x_2 \geqslant -6 \end{cases} \tag{2}$$

或

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \\ -3 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \geqslant \begin{bmatrix} 2 \\ -1 \\ -6 \end{bmatrix}$$

满足 $x_1 + x_2 \geqslant 2$ 的点位于由 $x_1 + x_2 = 2$ 给出的直线的一边．由图可知，这条线和其他的两条线

由系统(2)所产生，我们可以确定表示系统(2)的解的所有点的集合，它是图 10-1 中的一个三角形. 从图中容易看出，如果把系统(2)中所有的不等式反向，则所得的系统是**不相容的**. 而且，在系数矩阵中作某些小的改变，可以产生一个系统，它的解集是**无界的**.（例如，在第三个不等式中把-3变为-1.）这也粗略地表明解集是**凸的**，现在详细阐述这个事实.（关于凸集的理论，读者可查阅 6.9 节.）



682

图 10-1 系统(2)的解集

10.1.2 凸集和凸包

定义 1（凸集的定义） 一个线性空间中的一个子集 K 称为**凸的**，如果连接 K 的任何两个点的线段全部落在 K 中.

这个性质的代数表达式为

$$\left.\begin{array}{l} x, y \in K \\ 0 \leqslant \theta \leqslant 1 \end{array}\right\} \Rightarrow \theta x + (1-\theta) y \in K$$

关于凸集的某些基本事实包含于下面一些定理中. 术语**凸组合**表示点的线性组合，其中系数是非负的且它们的和等于 1.

定理 1（凸集定理） 若 K 是凸的，则 K 中的点的任何凸组合也属于 K.

证明 定理 1 用归纳法来证明. $m=1$ 的情形是显而易见的，$m=2$ 的情形由凸性的定义可知是正确的. 为了从 $m-1$ 的情形证明 m 的情形，我们记

$$\sum_{i=1}^{m} \lambda_i x^{(i)} = \lambda_m x^{(m)} + (1-\lambda_m) \sum_{i=1}^{m-1} \frac{\lambda_i}{1-\lambda_m} x^{(i)}$$

其中 $x^{(i)}$ 是 K 中的点，$\lambda_i \geqslant 0$ 且 $\sum_{i=1}^{m} \lambda_i = 1$. 因而系数 $\lambda_i/(1-\lambda_m)$ 是非负的且其和等于 1. ■

定理 2（凸集之交定理） 凸集族之交也是凸的.

证明 假设对一个指标集中的每个 α，K_α 是一个凸集. 如果 x，$y \in \bigcap K_\alpha$ 且 $0 \leqslant \theta \leqslant 1$，因为 K_α 是凸的，所以 $\theta x + (1-\theta) y \in K_\alpha$ 对每个 α 成立. 因此，$\theta x + (1-\theta) y \in \bigcap K_\alpha$. ■

下面是一些涉及系统(2)解集的定理.

定理 3(解集定理) 线性不等式系统的解集是一个凸集.

证明 单个线性不等式$a^{T}x\geqslant\beta$的解集是凸的，这可以通过下列计算证明

$$a^{T}(\theta x+(1-\theta)y)=\theta a^{T}x+(1-\theta)a^{T}y\geqslant\theta\beta+(1-\theta)\beta=\beta$$

其中x和y是这个解集中的元. 线性不等式系统的解集是单个不等式解集的交. 因为任何凸集族之交是凸的，所以这个集合也是凸的. ■

回顾6.9节，集合S的**凸包**是一切S中点的凸线性组合的集合，而且通常用co(s)表示. [683]

定理 4(凸包定理) 集合S的凸包是包含S的最小的凸集.

证明 设K是S的凸包,又设T是其他任意包含S的凸集,我们要证明$K\subseteq T$. K中一个典型的元素具有形式$x=\sum_{i=1}^{n}\lambda_i x_i$,其中$x_i\in S,\lambda_i\geqslant 0$,且$\sum_{i=1}^{n}\lambda_i=1$. 因为$T$包含$S$,所以显然有$x_i\in T$. 因为$T$是凸的,所以$x\in T$. 因为$x$是$K$的一个任意的元素,所以$K\subseteq T$. ■

定理 5(分离定理 I) 设X是$\mathbb{R}^n$中的一个闭凸集，如果p是一个不在X中的点，则对某个$v\neq 0$，我们有

$$\langle v,p\rangle<\inf_{x\in X}\langle v,x\rangle$$

证明 设S是一个中心在p的闭球，它具有足够大的半径保证S与X相交，因而$S\cap X$是紧的. 函数$x\mapsto\|x-p\|$是连续的，假定在$S\cap X$上它的极小值是ξ. 如果$x\in X$且$0<\theta<1$，则$y\equiv\theta x+(1-\theta)\xi\in K$，因此容易验证$\|y-p\|\geqslant\|\xi-p\|$. 于是，

$$\begin{aligned}\|\xi-p\|^2&\leqslant\|\theta x+(1-\theta)\xi-p\|^2\\&=\|\xi-p+\theta(x-\xi)\|^2\\&=\|\xi-p\|^2+2\theta\langle\xi-p,x-\xi\rangle+\theta^2\|x-\xi\|^2\end{aligned}$$

因此，

$$0\leqslant 2\langle\xi-p,x-\xi\rangle+\theta\|x-\xi\|^2$$

设θ收敛于0，我们得到

$$0\leqslant\langle\xi-p,x-\xi\rangle$$

设$v=\xi-p$，可记最后的不等式为

$$0\leqslant\langle v,x-p+p-\xi\rangle=\langle v,x-p-v\rangle=\langle v,x-p\rangle-\|v\|^2$$

因此，得到一个比我们断言的更强的不等式

$$0<\|v\|^2\leqslant\langle v,x\rangle-\langle v,p\rangle$$ ■

证明(另一种证明) 因为p不在X中，所以0不在平移集

$$X-p=\{x-p:x\in X\}$$ [684]

中. 应用6.9节的引理2于集合$X-p$. 通过考察该引理的证明，我们推断存在一个具有下列性质的向量$v\neq 0$，

$$\langle v,x-p\rangle\geqslant\langle v,v\rangle\qquad(x\in X)$$

由此得到

$$\langle v,p\rangle\leqslant\langle v,x\rangle-\|v\|^2\qquad(x\in X)$$ ■

上述定理在希尔伯特空间也成立，但是必须修正其证明．因为希尔伯特空间中的有界闭集不一定是紧的，所以点 ξ 的存在性现在不能根据 $S\cap X$ 的紧性得到．我们可以这样处理：设 $d=\inf\{\|x-p\|: x\in X\}$，选取一个序列 $x_i\in X$，使得 $\|x_i-p\|\to d$．由平行四边形定律，

$$\begin{aligned}\|x_i-x_j\|^2 &= \|(p-x_j)-(p-x_i)\|^2\\ &= 2\|p-x_j\|^2+2\|p-x_i\|^2-4\|p-(x_i+x_j)/2\|^2\\ &\leqslant 2\|p-x_j\|^2+2\|p-x_i\|^2-4d^2\to 0\end{aligned}$$

这表明序列 x_i 有柯西性质．因此，它收敛于点 ξ，并且容易看出 $\xi\in X$（因为 X 是闭的）而且 $\|x-\xi\|=d$（因为连续性）．证明的其余部分对希尔伯特空间来说不需改变．

分离定理 I 的适当形式在更一般的空间是成立的．现在列举一个几何上的应用例子．$\mathbb{R}^n$ 中的一个**闭半空间**定义为下列形式的集合

$$\{x:\langle a,x\rangle\geqslant\lambda\}$$

其中 $a\in\mathbb{R}^n$，$\lambda\in\mathbb{R}$ 且 $a\neq 0$．

定理 6（半空间定理） $\mathbb{R}^n$ 中每个闭凸集是一切包含它的闭半空间的交．

证明 设 X 是一个闭凸集．显然，X 包含于那些包含 X 的闭半空间的交之中．对于相反的结论，假设 p 是一个不在 X 中的点．由分离定理 I，存在一个向量 $v\neq 0$，使得

$$\langle v,p\rangle<\inf_{x\in X}\langle v,x\rangle$$

取

$$\lambda=\inf_{x\in X}\langle v,x\rangle$$

我们看出

$$X\subseteq\{x:\langle v,x\rangle\geqslant\lambda\}$$

685 但是这个半空间不包含 p． ■

定理 7（分离定理 II） 设 X 是闭凸集而 Y 是紧凸集（$\mathbb{R}^n$ 中），若 X 和 Y 不交，则存在一个向量 $v\in\mathbb{R}^n$，使得

$$\inf_{x\in X}\langle v,x\rangle>\sup_{y\in Y}\langle v,y\rangle$$

证明 首先我们指出集 $Z=X-Y$ 是闭的，设 $z_k\in Z$ 且假定 $z_k\to z$，问 $z\in Z$ 吗？我们记 $z_k=x_k-y_k$，$x_k\in X$ 且 $y_k\in Y$．由 Y 的紧性，存在一个收敛的子列 $y_{k_i}\to y\in Y$．于是 $z_{k_i}\to z$ 且 $x_{k_i}\to z+y$．因为 X 是闭的，所以 $z+y\in X$，因此，$z\in X-Y$．其次我们注意到 $0\notin x-y$，这是因为 $0=x-y$ $(x\in X,\ y\in Y)$ 将推出 X 和 Y 包含一个公共点．此外，快速的计算表明 Z 是凸的．所以可对 Z 应用分离定理 I，推出存在一个向量 v，使得

$$\langle v,0\rangle<\inf_{x\in X,y\in Y}\langle v,x-y\rangle$$

如果 ε 表示不等式中的下确界，对 $x\in X$ 和 $y\in Y$，我们有 $\langle v,x-y\rangle\geqslant\varepsilon$，这就立即导出定理陈述中所断言的不等式． ■

10.1.3 极值点

一个凸集 K 的**极值点**是一个点 $x\in K$，它不能写成 $x=\theta y+(1-\theta)z$，其中 $0<\theta<1$，$y\in K$，$z\in K$ 且 $y\neq z$．换句话说，它不是属于 K 的任意线段的**内点**．一个等价的定义是如果 $K\setminus\{x\}$ 是凸的，则 x 是凸集 K 的一个极值点．例如，立方体的顶点是立方体仅有的极值点，实心球的极值点是它所有的边界点．

下面的定理是 **Krein-Milman 定理的有限维形式**(该定理可参见 Royden[1968]).

定理 8(Krein-Milman 定理，有限维形式) *在 n 维空间中每一个紧凸集是它的极值点集合的凸包的闭包.*

证明 对 n 用归纳法证明. 若 $n=1$，则凸集是一个有界闭区间. 极值点是区间的端点，所以此时定理显然成立. 现在假定定理对维数小于 n 的情形成立，并设 K 是 n 维空间中的一个紧凸集. 设 E 是 K 的极值点的集合，并设 H 表示 E 的凸包. 要证明 $\overline{H}=K$($\overline{H}$ 表示 H 的闭包). 因为 K 是凸的且 $E\subseteq K$，所以有 $H\subseteq K$. 因为 K 是闭的，所以 $\overline{H}\subseteq K$. 我们必须证明这个后面的包含不是真包含. 假定 p 是 $K\setminus\overline{H}$ 的一个点，利用平移($x\mapsto x-p$)，我们可假定 $p=0$. 因为 $0\notin\overline{H}$，所以齐次不等式定理(6.9 节定理 3)推出存在一个向量 v，使得对一切 $u\in\overline{H}$ 有 $\langle v, u\rangle<0$. 我们取

686

$$c=\sup\{\langle v,x\rangle : x\in K\}$$

因为 $0\in K$，所以有 $c\geqslant 0$. 因为 K 是紧的，所以这个上确界可达到，这意味着集合

$$K'=\{x\in K:\langle v,x\rangle=c\}$$

非空. 因为集合 K'位于一个超平面上，所以 K'也是紧凸的且维数为 $n-1$. 由归纳假设，K'至少有一个极值点 z. 正如我们将证明的那样，z 事实上就是 K 的一个极值点，假设 $z=\theta z_1+(1-\theta)z_2$，$0<\theta<1$ 且 $z_i\in K$. 则

$$c=\langle v,z\rangle=\theta\langle v,z_1\rangle+(1-\theta)\langle v,z_2\rangle\leqslant\theta c+(1-\theta)c=c$$

因此，$\langle v, z_1\rangle=\langle v, z_2\rangle=c$ 且 $z_i\in K'$. 但是 z 是 K'的一个极值点，所以 $z_1=z_2$. 这就证明了 $z\in E$. 因此，$\langle v, z\rangle<0\leqslant c$：矛盾. ■

在最优化问题中极值点的重要性源自于在紧凸集上求线性函数极小问题，我们仅限于搜索极值点. 下面是一个正式的结果.

定理 9(极大/极小性质定理) *设 K 是$\mathbb{R}^n$ 中的一个紧凸集，且设 f 是$\mathbb{R}^n$ 上的一个线性泛函，则 f 在 K 上的极大值和极小值在 K 的极值点上达到.*

证明 设

$$c=\sup\{f(x) : x\in K\}$$

因为 f 连续且 K 是紧的，所以集合

$$K'=\{x\in K : f(x)=c\}$$

非空. 它也是紧凸的. 因此，由 Krein-Milman 定理，K'具有一个极值点 z. 根据 Krein-Milman 定理证明的常见理由，z 是 K 的一个极值点. 通过考虑 $-f$ 的极大值可得 f 在 K 上的极小值的证明. ■

定理 9 对任何局部凸线性拓扑空间中的紧凸集均成立，它对任何连续凸泛函也是正确的. Krein-Milman 定理对任何局部凸线性拓扑空间情况是正确的，在有限维情况下，它不一定要取闭包，例如，参见 Holmes[1972，第 82 页].

习题 10.1

1. 证明每个闭凸集是包含它的一切开半空间的交. 开半空间是一个形如$\{x:\langle a, x\rangle>\lambda\}$的集合.

687

2. 证明 $p\in\mathrm{co}(X)$当且仅当 $0\in\mathrm{co}(X-p)$.

3. 证明：若 S 和 T 是凸集，则 λS，$S+T$ 和 $S-T$ 是凸的.（集 $S+T$ 定义为一切和 $s+t$ 的集合，其中 $s\in S$，$t\in T$.）
4. 证明：若 L 是一个线性映射且 K 是一个凸集，则 $L(K)$是凸的.（集 $L(K)$定义为一切点 $L(x)$的集合，其中 $x\in K$.）
5. 证明：$\mathrm{co}(\lambda S)=\lambda\mathrm{co}(S)$，$\mathrm{co}(S+T)=\mathrm{co}(S)+\mathrm{co}(T)$.
6. 设 X 是希尔伯特空间中的一个闭凸集，证明：若 $p\notin X$，$\xi\in X$ 且 $\|p-\xi\|=\mathrm{dist}(p,X)$，则对一切 $x\in X$，$\langle p-\xi, x-\xi\rangle\leqslant 0$.
7. 设 U 是$\mathbb{R}^n$ 中的一个紧集，证明：若对 $u\in U$，线性不等式系统$\langle u, x\rangle>0$ 不相容，则它包含一个至多有$n+1$ 个不等式的不相容子系统.
8. 证明：若 K 是希尔伯特空间中的一个闭凸集，则对每个点 $p\notin K$，在 K 中存在唯一的最近的点 k.
9. （续）证明上题中定义的映射 $p\mapsto k$ 是非扩张的，因此，若(p_1, k_1)和(p_2, k_2)是映射对，则 $\|k_1-k_2\|\leqslant\|p_1-p_2\|$.
10. 一个有界集能有一个凸补集吗？
11. 证明：若 X_i 是凸集，则 $\mathrm{co}(X_1\cup\cdots\cup X_k)$ 是一切凸组合 $\sum_{i=1}^{k}\theta_i x_i$ 的集合，其中 $x_i\in X_i$.
12. 证明凸集的闭包是凸的.
13. 设 K 是一个凸集，p 是 K 的内点，而 q 是 K 的任意点，证明：若 $0<\theta<1$，则 $\theta p+(1-\theta)q$ 是 K 的一个内点.
14. 证明：对 $u\in U$，若不等式系统$\langle u, x\rangle>0$ 是相容的，则对 $u\in\mathrm{co}(U)$，系统$\langle u, x\rangle>0$ 也是相容的.
15. 证明：若 S 和 T 是紧凸的，则 $\mathrm{co}(S\cup T)$也是紧凸的.
16. 证明：若 X 是$\mathbb{R}^n$ 中的一个有界集，则对任何 p，
$$\sup\{\|p-x\|:x\in X\}=\sup\{\|p-x\|:x\in\mathrm{co}(X)\}$$
17. 证明：若 U 是紧凸集且系统$\langle u, x\rangle>0$ 不相容$(u\in U)$，则当 u 位于 U 的极值点集合时，系统$\langle u, x\rangle>0$ 也不相容.
18. 证明：平面中任何凸集是顶点位于给定集合中一切三角形的并集.
19. 设 X 是至少包含$\mathbb{R}^n$ 中 $n+2$ 个点的有限集，证明它可以写成 $X=X_1\cup X_2$，其中 $\mathrm{co}(X_1)\cap\mathrm{co}(X_2)=\varnothing$.
20. 证明$\mathbb{R}^n$ 中的一个开集的凸包是开的.
21. 举例说明一个闭集的凸包不一定是闭的.
22. 集合 X 可能位于无限维的线性空间中，设 H_n 是可以写成 $\sum_{i=1}^{n}\theta_i x_i$ 的一切点的集合，$\theta_i\geqslant 0$，$\sum_{i=1}^{n}\theta_i=1$，$x_i\in X$. 证明 $\bigcup_{n=1}^{\infty}H_n$ 是凸的. 证明这个集合是 $\mathrm{co}(X)$.
23. 在$\mathbb{R}^n$ 中，设连接 x 和 y 的线段用$\overline{xy}$表示，对给定的集合 X_0，用
$$X_{k+1}=\bigcup\{\overline{xy}:x\in X_k, y\in X_k\}$$
来定义 X_1，X_2，…，X_k，证明：$X_{2^n+1}=\mathrm{co}(X_0)$.

688

10.2 线性不等式

设 X 是实数域上的一个向量空间，**线性泛函**是 X 到$\mathbb{R}$ 中的一个线性映射，**线性不等式**是一个命题 $f(x)\geqslant\alpha$，其中 f 是一个线性泛函. 这类单个的不等式具有一个**半空间**$\{x: f(x)\geqslant\alpha\}$作为它的解集. 更加有趣的是**线性不等式系统** $f_i(x)\geqslant\alpha_i$，$i\in I$，其中 I 是某个指标集，不一

定有限.

10.2.1 齐次方程组

我们从与齐次方程组有关的线性代数定理开始，然后转到一个类似的与不等式有关的定理. 用 f^0 表示**泛函 f 的零空间**：

$$f^0 = \{x \in X : f(x) = 0\}$$

并用 $\mathcal{L}(f_1, f_2, \cdots, f_m)$ 表示集合 $\{f_1, f_2, \cdots, f_m\}$ 的**线性生成**.

定理 1(线性泛函定理) 对一组线性泛函，下列性质等价：

1. $f^0 \supset \bigcap_{i=1}^{m} f_i^0$.
2. $f \in \mathcal{L}(f_1, f_2, \cdots, f_m)$.

证明 性质 2⇒ 性质 1 是容易的，事实上，如果 $f = \sum_{i=1}^{m} \lambda_i f_i$ 且 $x \in \bigcap_{i=1}^{m} f_i^0$，则对一切 i 有 $f_i(x) = 0$，由此显然可得 $f(x) = 0$.

反之，我们对 m 用归纳法，设 $m=1$ 且假设 $f^0 \supset f_1^0$，若 $f_1 = 0$，则 $f_1^0 = X$. 所以，$f^0 = X$ 且 $f=0$. 因此，在此情形下 $f \in \mathcal{L}(f_1)$，若 $f_1 \neq 0$，取一个点 y 使 $f_1(y) = 1$，且设 x 是 X 中的任意点. 我们有 $f_1[x - f_1(x)y] = f_1(x) - f_1(x)f_1(y) = 0$，故 $x_1 - f_1(x)y \in f_1^0$. 由假设，$x - f_1(x)y \in f^0$. 因此 $f(x) - f_1(x)f(y) = 0$ 或 $f = f(y)f_1$. 于是，$f \in \mathcal{L}(f_1)$.

对归纳步，假设定理对整数 m 成立. 假设

$$f^0 \supset \bigcap_{i=1}^{m+1} f_i^0$$

设 $Y = f_{m+1}^0$，这是 X 的一个子空间. 用记号 $f \mid Y$ 表示 f 在 Y 上的限制，我们有

$$(f \mid Y)^0 \supset \bigcap_{i=1}^{m} (f_i \mid Y)$$

689

由归纳假设，对适当的 λ_i，

$$f \mid Y = \sum_{i=1}^{m} \lambda_i f_i \mid Y$$

现在两个等价式是 $(f - \sum_{i=1}^{m} \lambda_i f_i) \mid Y = 0$ 和 $(f - \sum_{i=1}^{m} \lambda_i f_i)^0 \supset f_{m+1}^0$. 利用本定理 $m=1$ 的情形，我们得出，对某个 λ_{m+1}，

$$f - \sum_{i=1}^{m} \lambda_i f_i = \lambda_{m+1} f_{m+1}$$ ■

10.2.2 线性不等式

现在讨论关于线性不等式的一个完全类似的定理. 代替泛函的零空间，我们使用**半空间**：

$$f^+ = \{x \in X : f(x) \geqslant 0\}$$

代替通常的 $f_1, f_2, \cdots, f_m$ 的线性生成，我们考虑它们生成的**锥**：

$$\mathcal{C}\{f_1, f_2, \ldots, f_m\} = \left\{ \sum_{i=1}^{m} \lambda_i f_i : \lambda_i \geqslant 0 \right\}$$

定理 2(福科什定理，1902) 对线性泛函 f 和 f_i，下列性质是等价的：

1. $f^+ \supset \bigcap_{i=1}^{m} f_i^+$.

2. $f \in \mathcal{C}(f_1, f_2, \cdots, f_m)$.

证明 如果性质 2 成立，则性质 1 容易推出. 事实上，假设

$$f = \sum_{i=1}^{m} \lambda_i f_i$$

其中 $\lambda_i \geqslant 0$. 若 $x \in \bigcap_{i=1}^{m} f_i^+$，则对一切 i，$f_i(x) \geqslant 0$，并且由此显然可得 $f(x) \geqslant 0$.

反之，我们给出在 $X = \mathbb{R}^n$ 的假定下的证明. 对一个适当的 v，每一个线性泛函具有形式 $f(x) = \langle v, x \rangle$，假定 $f \notin C$，其中 C 是锥 $\mathcal{C}(f_1, f_2, \cdots, f_m)$. 由分离定理 I，存在一个向量 v，使得

[690]

$$\langle v, f \rangle < \inf_{u \in C} \langle v, u \rangle$$

设

$$k = \inf_{u \in C} \langle v, u \rangle$$

因为 $0 \in C$，所以有 $k \leqslant 0$. 若 $k < 0$，则选择 $u \in C$，使得 $k \leqslant \langle v, u \rangle < 0$. 对一切正的 t，有 $tu \in C$. 若 t 充分大，则有 $\langle v, tu \rangle < k$，这是一个矛盾. 因此，$k = 0$ 且对 $u \in C$，

$$\langle v, f \rangle < 0 \leqslant \langle v, u \rangle$$

由此可得 $f(v) < 0 \leqslant f_i(v)$，这表明性质(1)不真. ■

福科什定理具有下列矩阵-向量形式.

定理 3(矩阵-向量福科什定理) 矩阵 A 和向量 c 的下列性质等价：

1. 对一切 x，若 $Ax \geqslant 0$，则 $c^{\mathrm{T}} x \geqslant 0$.

2. 对某个 y，$y \geqslant 0$ 且 $c = A^{\mathrm{T}} y$.

10.2.3 相容系统和不相容系统

下面我们介绍与定理 3 类似的一个非齐次的定理，定理中我们使用下列术语：一个不等式系统 $f_i(x) \geqslant \alpha_i$ 被称为是另一个系统 $g_i(x) \geqslant \beta_i$ 的后承，如果第二个系统的每个解满足第一个系统. 因而

$$\{x : g_i(x) \geqslant \beta_i, \text{对一切 } i\} \subseteq \{x : f_i(x) \geqslant \alpha_i, \text{对一切 } i\}$$

定理 4(非齐次的福科什定理) 若线性不等式 $f(x) \geqslant \alpha$ 是线性不等式相容系统

$$g_i(x) \geqslant \beta_i \qquad (1 \leqslant i \leqslant n)$$

的一个后承，则对适当的 $\theta_i \geqslant 0$，我们有

$$f = \sum_{i=1}^{n} \theta_i g_i \quad \text{且} \quad \sum_{i=1}^{n} \theta_i \beta_i \geqslant \alpha$$

证明 考察系统

$$g_i(x) - \lambda\beta_i \geqslant 0 \qquad (\lambda > 0) \qquad (1 \leqslant i \leqslant n) \tag{1}$$

若对(x, λ)是系统(1)的解，则 $g_i(x) \geqslant \lambda\beta_i$ 且 $g_i(x/\lambda) \geqslant \beta_i$. 由定理的假设，$f(x/\lambda) \geqslant \alpha$. 因此

[691]

$$f(x) - \lambda\alpha \geqslant 0 \tag{2}$$

这表明不等式(2)是系统(1)的一个后承，现在考察系统

$$g_i(x)-\lambda\beta_i \geqslant 0 \qquad (\lambda \geqslant 0) \qquad (1 \leqslant i \leqslant n) \tag{3}$$

如果不等式(2)是系统(3)的一个后承，则应用福科什定理的齐次形式，我们得到

$$(f;-\alpha)=\sum_{i=1}^{n}\theta_i(g_i;-\beta_i)+\theta_0(0;1) \qquad (\theta_i \geqslant 0) \qquad (1 \leqslant i \leqslant n)$$

其中$(f_j;\ -\alpha)$简单地表示一个向量对．这个结论可写成下列形式

$$f=\sum_{i=1}^{n}\theta_i g_i \qquad \alpha=\sum_{i=1}^{n}\theta_i\beta_i-\theta_0 \leqslant \sum_{i=1}^{n}\theta_i\beta_i$$

这就是要证明的论断，然而，我们并没有完成证明，因为可能发生(2)不是(3)的后承的情况，虽然我们已指出它是(1)的一个后承．此时，(3)存在一个既不是(1)也不是(2)的解．这样一个解对$(u;\ \lambda)$必须有$\lambda=0$，因为否则的话它是(1)的解，从而也是(2)的解．于是，$g_i(u)\geqslant 0>f(u)$．由定理的假设，存在一个向量v，使得$g_i(v)\geqslant\beta_i$．选取一个正数λ使得$f(u+\lambda v)<\lambda\alpha$，这是可能的．因为当$\lambda\downarrow 0$时，上式的右边逼近于0而左边逼近于一个负数$f(u)$．因为$f(u/\lambda+v)<\alpha$而$g_i(u/\lambda+v)\geqslant g_i(v)\geqslant\beta_i$，因而这与假设矛盾． ■

10.2.4 矩阵-向量形式

定理4的矩阵-向量形式如下．

定理5(矩阵-向量非齐次福科什定理) 若系统

$$Ax \geqslant b$$

相容，而系统

$$Ax \geqslant b \qquad c^{\mathrm{T}}x<\alpha$$

不相容，则系统

$$A^{\mathrm{T}}y=c \qquad y^{\mathrm{T}}b \geqslant \alpha \qquad y \geqslant 0$$

相容．

692

定理6(不相容系统第一定理) 若系统

$$Ax=b \qquad x \geqslant 0 \tag{4}$$

不相容，则系统

$$A^{\mathrm{T}}y \geqslant 0 \qquad b^{\mathrm{T}}y \leqslant 0 \tag{5}$$

相容．

证明 若系统(4)不相容，则

$$b \notin K \equiv \{Ax : x \geqslant 0\}$$

因为K是闭凸的，应用10.1节分离定理I，存在向量y，使得

$$\langle y,b\rangle < \inf_{x\geqslant 0}\langle y,Ax\rangle$$

因为在这个计算中x可能为0，所以我们有$\langle y,\ b\rangle<0$．留待我们证明的是$A^{\mathrm{T}}y\geqslant 0$，如果它不成立，则对某个指标$\alpha$，$(A^{\mathrm{T}}y)_\alpha<0$．设$x$是坐标为$x_j=\lambda\delta_{\alpha j}$的向量．于是

$$\langle y,Ax\rangle=\sum_{j=1}^{n}(A^{\mathrm{T}}y)_j x_j=\lambda(A^{\mathrm{T}}y)_\alpha \to -\infty(\text{当 }\lambda\to+\infty)$$

因而对适当的λ，我们必有$\langle y,\ Ax\rangle<\langle y,\ b\rangle$，这是一个矛盾．因此，$y$是系统(5)的解． ■

定理 7(不相容系统第二定理) 若系统

$$Ax \leqslant b \qquad x \geqslant 0 \tag{6}$$

不相容，则系统

$$A^{T}y \geqslant 0 \qquad b^{T}y < 0 \qquad y \geqslant 0 \tag{7}$$

相容.

证明 若系统(6)不相容，则改写(6)为

$$Ax + z = b \qquad x \geqslant 0 \qquad z \geqslant 0 \tag{8}$$

我们把系统(8)写成下列形式

693

$$[A \quad I]\begin{bmatrix} x \\ z \end{bmatrix} = b \qquad \begin{bmatrix} x \\ z \end{bmatrix} \geqslant 0 \tag{9}$$

由定理 6，系统

$$\begin{bmatrix} A^{T} \\ I \end{bmatrix} y \geqslant 0 \qquad b^{T}y < 0 \tag{10}$$

相容. 这是系统(7)的另一种形式. ■

习题 10.2

1. 设 A 为 $m \times n$ 矩阵，证明下列这些系统中的一个相容但不会两个都相容：
   a. $Ax=0$，$x \geqslant 0$，$x \neq 0$
   b. $A^{T}y>0$
2. 设 A 为 $m \times n$ 矩阵，证明下列这些系统中的一个相容但不会两个都相容：
   a. $Ax \leqslant 0$，$x \geqslant 0$，$x \neq 0$
   b. $A^{T}y>0$，$y>0$
   （注：这称为 **Ville 定理**，1938.）
3. 定义 $P_n = \{x \in \mathbb{R}^n : x \geqslant 0 \text{ 且 } \sum_{i=1}^{n} x_i = 1\}$，设 A 为任意的 $m \times n$ 矩阵，证明：要么对某个 $x \in P_n$，$Ax \geqslant 0$，要么对某个 $y \in P_m$，$A^{T}y \leqslant 0$ 成立.
4. (续)利用上题的记号，证明对任意的 $m \times n$ 矩阵 A 有
   $$\max_{x \in P_n} \min_{y \in P_m} y^{T}Ax \leqslant \min_{y \in P_m} \max_{x \in P_n} y^{T}Ax$$
   提示：从 $\min_y y^{T}Ax \leqslant \max_x y^{T}Ax$ 着手.
5. 如果 U 是全部元素为 1 的 $m \times n$ 矩阵，证明对一切 $x \in P_n$ 和一切 $y \in P_m$ 有
   $$y^{T}(A-\lambda U)x = y^{T}Ax - \lambda$$
6. (**博弈论的极小-极大定理**)证明
   $$\max_{x \in P_n} \min_{y \in P_m} y^{T}Ax = \min_{y \in P_m} \max_{x \in P_n} y^{T}Ax$$
   提示：若上面习题 10.2.4 中的不等式是严格的，则设 λ 为两个量中间的数. 利用习题 10.2.5，对矩阵 $A-\lambda U$ 应用习题 10.2.3.
7. 证明：若不等式 $Ax \geqslant b$ 无解，则对某个 y，下式成立：$y \geqslant 0$，$A^{T}y=0$ 且 $b^{T}y=1$.
8. 证明：若不等式 $Ax \geqslant b$ 没有非负解，则不等式 $A^{T}y \leqslant 0$，$b^{T}y>0$ 有非负解.
9. 证明：若系统 $Ax=0$，$x \geqslant 0$，$x \neq 0$ 不相容，则系统 $A^{T}y<0$ 相容.

10. 证明：若系统 $Ax\geqslant 0$，$x\geqslant 0$，$x\neq 0$ 不相容，则系统 $A^{\mathrm{T}}y<0$，$y\geqslant 0$ 相容.

11. 证明：若系统 $Ax=0$，$x>0$ 不相容，则系统 $Ax\leqslant 0$ 相容.

12. 证明：若 A 是一个 $n\times(n+1)$ 矩阵，则系统 $Ax\geqslant 0$，$x\neq 0$ 相容.

13. 证明：对任意的 $m\times n$ 矩阵 A，系统 $Ax>0$ 相容当且仅当系统 $A^{\mathrm{T}}y=0$，$y>0$ 不相容. 694

14. 求集合

$$K=\{x\in\mathbb{R}^n : Ax\leqslant b\}$$

有界的充分必要条件.

15. 证明下列系统中的一个相容但不会两个都相容：

a. $Ax=b$

b. $A^{\mathrm{T}}y=0$，$b^{\mathrm{T}}y=1$

10.3 线性规划

术语**线性规划**与计算机的程序设计无关，但是与商业或经济计划的规划有关. 一个明确的技术上的意思是指：在$\mathbb{R}^n$ 中的一个凸多面体集上求 n 个实变量的一个线性函数的最大值. 对这样的问题，我们采用下列标准形式.

LP 问题 1(线性规划问题：第一标准形式) 设 $c\in\mathbb{R}^n$，$b\in\mathbb{R}^m$，$A\in\mathbb{R}^{m\times n}$，求 $c^{\mathrm{T}}x$ 的最大值，它服从于约束 $x\in\mathbb{R}^n$，$Ax\leqslant b$ 和 $x\geqslant 0$.

我们提醒读者，若 $x=(x_1, x_2, \cdots, x_n)^{\mathrm{T}}$，则向量不等式 $x\geqslant 0$ 表示对一切 $i\in\{1, 2, \cdots, n\}$，$x_i\geqslant 0$. 类似地，不等式

$$Ax\leqslant b$$

表示对一切 $i\in\{1, 2, \cdots, m\}$，

$$\sum_{j=1}^{n}a_{ij}x_j\leqslant b_i$$

下面描述一些常用的术语. 在我们的问题中**可行集**是集合

$$K=\{x\in\mathbb{R}^n : Ax\leqslant b, x\geqslant 0\}$$

问题的**值**是数

$$v=\sup\{c^{\mathrm{T}}x : x\in K\}$$

可行点是 K 的任意元素. **解**或**最优可行点**是使 $c^{\mathrm{T}}x=v$ 的任意 $x\in K$，函数 $x\mapsto c^{\mathrm{T}}x=\sum_{j=1}^{n}c_jx_j$ 是**目标函数**. 因为问题是由数据 A,b 和 c 完全确定的，所以我们把它称为线性规划问题(A,b,c). 695

10.3.1 转换问题的方法

几乎任何涉及变量服从于线性不等式的线性函数的最优化问题都可以转化为线性规划格式. 做这个转化通常需要下列一个或几个观念：

1. 若希望极小化 $c^{\mathrm{T}}x$，则这与极大化 $-c^{\mathrm{T}}x$ 相同.
2. 任何形如 $a^{\mathrm{T}}x\geqslant\beta$ 的约束等价于 $-a^{\mathrm{T}}x\leqslant-\beta$.
3. 任何形如 $a^{\mathrm{T}}x=\beta$ 的约束等价于 $a^{\mathrm{T}}x\leqslant\beta$，$-a^{\mathrm{T}}x\leqslant-\beta$.
4. 任何形如 $|a^{\mathrm{T}}x|\leqslant\beta$ 的约束等价于 $a^{\mathrm{T}}x\leqslant\beta$，$-a^{\mathrm{T}}x\leqslant\beta$.
5. 若目标函数包含一个附加的常数，则它对解没有影响. 因此，$c^{\mathrm{T}}x+\beta$ 的极大值如 $c^{\mathrm{T}}x$

的极大值那样在同样的点上存在.

6. 如果给定的问题不需要变量 x_j 是非负的，则可用两个需要非负的变量的差替代 x_j，例如，$x_j=u_j-v_j$.

例 1 把下列线性规划问题转换成标准形式

Min：$7x_1-x_2+x_3-4$

$$\text{约束：}\begin{cases}x_1+x_2-x_3\geqslant 2\\3x_1+4x_2+x_3=6\\|x_1-2x_2+3x_3|\leqslant 5\\x_1\geqslant 0,\ x_2\leqslant 0\end{cases}$$

解 设 $u_1=x_1$，$u_2=-x_2$，$u_3-u_4=x_3$. 我们有

Max：$-7u_1-u_2-u_3+u_4$

$$\text{约束：}\begin{cases}-u_1+u_2+u_3-u_4\leqslant -2\\3u_1-4u_2+u_3-u_4\leqslant 6\\-3u_1+4u_2-u_3+u_4\leqslant -6\\u_1+2u_2+3u_3-3u_4\leqslant 5\\-u_1-2u_2-3u_3+3u_4\leqslant 5\\u_1\geqslant 0,\ u_2\geqslant 0,\ u_3\geqslant 0,\ u_4\geqslant 0\end{cases}$$

■

一个给定的线性规划问题(A, b, c)可能有解或者可能无解. 首先，可行集 K 可能是空的，因而无解. 若可行集非空且无界，可能发生目标函数在 K 上无上界. $v=+\infty$时，无解. 若 K 是空的，则 $v=-\infty$，无解. 若 K 非空且有界，则至少存在一个解. 这是 K 为紧的(在$\mathbb{R}^n$上有界闭的)这个事实的一个推论，因而目标函数(它是连续的)在 K 上达到它的上确界.

696

10.3.2 对偶问题

对任何线性规划问题(A, b, c)，我们可以联系另一个问题$(-A^{\mathrm{T}}, -c, -b)$，这个问题称为原问题的**对偶**. 例如，问题

Max：$3x_1-2x_2$

$$\text{约束：}\begin{cases}7x_1+x_2\leqslant 18\\-3x_1+5x_2\leqslant 25\\6x_1-x_2\leqslant 13\\x_1\geqslant 0,\ x_2\geqslant 0\end{cases}$$

有下列对偶问题：

Max：$-18y_1-25y_2-13y_3$

$$\text{约束：}\begin{cases}-7y_1+3y_2-6y_3\leqslant -3\\-y_1-5y_2+y_3\leqslant 2\\y_1\geqslant 0,\ y_2\geqslant 0,\ y_3\geqslant 0\end{cases}$$

一个线性规划问题和它的对偶之间的关系是**对偶性理论**的主题. 现在将讨论其某些显著的结果.

定理 1（第一线性规划和对偶问题定理） 若 x 是线性规划问题(A, b, c)的一个可行点，而 y 是对偶问题$(-A^{T}, -c, -b)$的一个可行点，则

$$c^{T}x \leqslant y^{T}Ax \leqslant b^{T}y$$

若这里出现等式，则 x 和 y 是它们各自问题的解.

证明 点 x 和 y 满足

$$x \geqslant 0 \qquad Ax \leqslant b \qquad y \geqslant 0 \qquad -A^{T}y \leqslant -c$$

由此可得

$$c^{T}x \leqslant (A^{T}y)^{T}x = y^{T}Ax \leqslant y^{T}b = b^{T}y$$

所以这两个问题的值 v_1 和 v_2 必须满足

$$c^{T}x \leqslant v_1 \leqslant b^{T}y$$
$$-b^{T}y \leqslant v_2 \leqslant -c^{T}x$$

若 $c^{T}x=b^{T}y$，则显然有 $c^{T}x=v_1=b^{T}y=-v_2$. ■ [697]

定理 1 通常可用于估计一个线性规划问题的值 v_1. 如果已知一个可行点 x，且如果已知对偶问题的一个可行点 y，则不等式 $c^{T}x \leqslant v_1 \leqslant b^{T}y$ 确定一个包含 v_1 的区间.

定理 2（第二线性规划和对偶问题定理） 若一个线性规划问题及其对偶有可行点，则两个问题同时有解，而且它们的值互为负的.

证明 由定理 1 可以证明存在 x 和 y 使得

$$x \geqslant 0 \qquad Ax \leqslant b \qquad y \geqslant 0 \qquad -A^{T}y \leqslant -c \qquad c^{T}x \geqslant b^{T}y$$

确实，这样的对(x, y)给出原问题的解 x 和它的对偶解 y. 我们的任务是证明下列线性不等式系统相容：

$$\begin{bmatrix} A & 0 \\ 0 & -A^{T} \\ -c^{T} & b^{T} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \leqslant \begin{bmatrix} b \\ -c \\ 0 \end{bmatrix} \qquad \begin{bmatrix} x \\ y \end{bmatrix} \geqslant \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

假定这个系统不相容并试图导出一个矛盾. 由 10.2 节中的定理 7，下列系统相容：

$$\begin{cases} \begin{bmatrix} A^{T} & 0 & -c \\ 0 & -A & b \end{bmatrix} \begin{bmatrix} u \\ v \\ \lambda \end{bmatrix} \geqslant \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ \begin{bmatrix} b^{T} & -c^{T} & 0 \end{bmatrix} \begin{bmatrix} u \\ v \\ \lambda \end{bmatrix} < 0 \\ \begin{bmatrix} u \\ v \\ \lambda \end{bmatrix} \geqslant 0 \end{cases}$$

这里 u 和 v 是向量而 λ 是一个常数，设(u, v, λ)满足这个系统，则

$$\begin{cases} A^{T}u - \lambda c \geqslant 0 & -Av + \lambda b \geqslant 0 & b^{T}u - c^{T}v < 0 \\ u \geqslant 0 & v \geqslant 0 & \lambda \geqslant 0 \end{cases}$$

首先假定$\lambda>0$，则$\lambda^{-1}v$对问题(A, b, c)可行，并且$\lambda^{-1}u$对对偶问题$(-A^T, -c, -b)$可行，因此，由定理1，我们有$c^T(\lambda^{-1}v)\leqslant b^T(\lambda^{-1}u)$以及$b^Tu-c^Tv\geqslant 0$，矛盾.

若$\lambda=0$，则$A^Tu\geqslant 0\geqslant Av$. 取$x$对原问题可行以及取$y$对对偶问题可行，我们得出一个与前面矛盾的不等式：

698

$$c^Tv\leqslant(A^Ty)^Tv=y^TAv\leqslant 0\leqslant(A^Tu)^Tx=u^T(Ax)\leqslant u^Tb=b^Tu$$ ■

定理3(*第三线性规划和对偶问题定理*) *若线性规划有解或它的对偶之一有解，则另一个有解.*

证明 因为对偶问题的对偶是原问题，所以只要证明这个定理的一种情形. 假设对偶问题$(-A^T, -c, -b)$有一个解y_0，则不等式系统

$$-A^Ty\leqslant -c \qquad y\geqslant 0 \qquad -b^Ty\geqslant -b^Ty_0$$

不相容. 我们省略第三个不等式的相应的系统当然相容，记不相容系统为下列形式

$$\begin{bmatrix}A^T\\I\end{bmatrix}y\geqslant\begin{bmatrix}c\\0\end{bmatrix} \qquad b^Ty\leqslant b^Ty_0$$

现在利用10.2节非齐次福科什定理，它推出系统

$$[A\quad I]\begin{bmatrix}x\\u\end{bmatrix}=b \qquad [x^T\quad u^T]\begin{bmatrix}c\\0\end{bmatrix}\geqslant b^Ty_0 \qquad \begin{bmatrix}x\\u\end{bmatrix}\geqslant\begin{bmatrix}0\\0\end{bmatrix}$$

的相容性，因而

$$Ax+u=b \qquad x^Tc\geqslant b^Ty_0 \qquad x\geqslant 0 \qquad u\geqslant 0$$

由此可得

$$Ax\leqslant b \qquad c^Tx\geqslant b^Ty_0 \qquad x\geqslant 0$$

因此，由定理1知x是原问题的解. ■

定理4(*可行点定理*) *设x和y分别是线性规划及其对偶的可行点，这些点是它们各自问题的解当且仅当对每个指标i，$(Ax)_i=b_i$使得$y_i>0$，并且对每个指标i，$(A^Ty)_i=c_i$使得$x_i>0$.*

证明 若x和y是解，则由定理1和2，

$$y^Tb=b^Ty=y^TAx=c^Tx=x^Tc$$

这就得到等式$y^T(b-Ax)=0$. 因为$y\geqslant 0$且$b-Ax\geqslant 0$，所以推出对每个i，$y_i(b_i-(Ax)_i)=0$. 因此，每当i是一个使$y_i>0$的指标时$(Ax)_i=b_i$，由对称的论断可得到另一个条件. 反之，对一切i，假设$y_i(b_i-(Ax)_i)=0$且$x_i(c_i-(A^Ty)_i)=0$，则

$$b^Ty=y^Tb=y^TAx=x^TA^Ty=x^Tc=c^Tx$$

699

由定理1知，x和y是它们各自问题的解. ■

习题10.3

1. 转换下列问题为标准的线性规划和对偶线性规划形式.

   a. Min：$3x_1+x_2-5x_3+2$

   $$约束：\begin{cases}x_1\geqslant x_2\\x_2\leqslant 0\\-x_1+4x_3\geqslant 0\\x_1+x_2+x_3=0\end{cases}$$

b. Min: $|x_1+x_2+x_3|$

约束: $\begin{cases} x_1-x_2=5 \\ x_2-x_3=7 \\ x_1\leqslant 0,\ x_3\geqslant 2 \end{cases}$

c. Min: $|x_1|-|x_2|$

约束: $\begin{cases} x_1+x_2=5 \\ 2x_1+3x_2-x_3\leqslant 0 \\ x_3\geqslant 4 \end{cases}$

2. 如果每一个可行点是线性规划问题(A, b, c)的一个解，你对线性规划问题(A, b, c)能证明什么?

10.4 单纯形法

利用标准的技巧，一个线性规划问题可以赋于下列形式:

LP 问题 2(线性规划问题: 第二标准形式) 求 $c^{\mathrm{T}}x$ 的极大值服从于 $Ax=b$, $x\geqslant 0$. 这里 $x\in\mathbb{R}^n$, $c\in\mathbb{R}^n$, $b\in\mathbb{R}^m$, $A\in\mathbb{R}^{m\times n}$.

10.4.1 基本概念

回忆表达式 $c^{\mathrm{T}}x$ 定义**目标函数**，我们注意到，若一个线性规划问题具有形如

$$Ax\leqslant b$$

的约束，则通过引进一个向量 $u\geqslant 0$ (其分量称为**松弛变量**)，我们可得

$$Ax+u=b$$

从而得到上面给出的第二标准形式. [700]

对第二标准形式问题，**可行集**是

$$K=\{x\in\mathbb{R}^n : Ax=b, x\geqslant 0\}$$

对任意的 $x\in K$，我们用

$$I(x)=\{i : 1\leqslant i\leqslant n \text{ 且 } x_i>0\}$$

定义一个 x 的分量是正的指标集. 矩阵 A 的列用 A_1, A_2, …, A_n 表示. 等式

$$Ax=b$$

变成

$$\sum_{i=1}^{n}x_iA_i=b$$

现在我们叙述一个重要的定理.

定理 1(线性规划性质定理) 设 $x\in K$，则下列 x 的性质等价:

1. x 是 K 的极值点.
2. $\{A_i : i\in I(x)\}$是线性无关的.

证明 假设性质 2 成立，我们将证明性质 1 成立. 设 $x=\theta u+(1-\theta)v$，其中 $u\in K$, $v\in K$ 且 $0<\theta<1$，对每个 $i\notin I(x)$，我们有

$$0=x_i=\theta u_i+(1-\theta)v_i$$

记得 $u_i\geqslant 0$ 和 $v_i\geqslant 0$，所以我们推出对于 $i\notin I(x)$，$u_i=v_i=0$. 于是有

$$0 = Au - Av = \sum_{i=1}^{n}(u_i - v_i)A_i = \sum_{i\in I(x)}(u_i - v_i)A_i$$

由性质 2 推出 $u_i = v_i$，因此 x 是 K 的一个极值点.

现设性质 1 成立，若

$$\sum_{i\in I(x)} w_i A_i = 0$$

则对 $i \notin I(x)$ 取 $w_i = 0$，显然地

[701]

$$\sum_{i\in I(x)}(x_i \pm \lambda w_i)A_i = b$$

因为当 $i\in I(x)$ 时，$x_i > 0$，所以可取 $\lambda \neq 0$ 并且如此之小，使得当 $i\in I(x)$ 时 $x_i + \lambda w_i > 0$ 且 $x_i - \lambda w_i > 0$. 因此，$u = x + \lambda w$ 和 $v = x - \lambda w$ 都是可行点. 因为 $x = \frac{1}{2}(u+v)$ 和 x 是 K 的极值点，所以我们得到 $u = v$ 以及 $w = 0$. 这就证得性质 2 成立. ■

推论 1(有限极值点推论) *可行集 K 只可能有有限个极值点.*

证明 设 E 是 K 的极值点集，对每个 $x\in E$，$I(x) \subseteqq \{1, 2, \cdots, n\}$，故 $I: E \to 2^{\{1,2,\cdots,n\}}$. (记号 2^S 表示**集合 S 的所有子集族**). 因而定义在 E 上的映射 I 是一一对应的. 为证实这点，设 $x, y\in E$ 且 $x\neq y$，则

$$b = Ax = \sum_{i=1}^{n} x_i A_i = \sum_{i\in I(x)} x_i A_i$$

$$b = Ay = \sum_{i=1}^{n} y_i A_i = \sum_{i\in I(y)} y_i A_i$$

若 $I(x) = I(y)$，则这与集合 $\{A_i : i\in I(x)\}$ 的线性无关性矛盾. E 单射到 $2^{\{1,2,\cdots,n\}}$ 中的映射表明 E 的元素个数不会超过 2^n. ■

Dantzig[1948]的单纯形法由两部分组成，第一部分求 K 的一个初始的极值点 $x^{(1)}$，第二部分从 $x^{(1)}$ 出发，生成一个有限的极值点序列，使得目标函数随着每一个生成的点递增. 用 $x^{(1)}$，$x^{(2)}$，…表示这个序列，由此可得 $c^{\mathrm{T}}x^{(1)} < c^{\mathrm{T}}x^{(2)} < \cdots$. 若问题无解，则在算法的进程中发现这个事实；若线性规划问题有解，则在算法的有限步骤，产生一个极值点 $x^{(k)}$，这是一个解，即它使目标函数达到最大值.

10.4.2 抽象形式

下面我们描述算法第 2 部分的抽象形式，我们采用下面的假设.

假设 1(非退化假设) *可行集 K 的每一个极值点 x 恰有 m 个正分量.*

假设 x 是 K 的一个极值点，由定理 1，集合 $\{A_i : i\in I(x)\}$ 线性无关，因此，由非退化假设知，它是 $\mathbb{R}^m$ 的一个基(由此原因，我们发现在文献中这样的 x 称为**基本可行点**). 一定存在

[702]

系数 D_{ij}，使得

$$A_j = \sum_{i\in I(x)} D_{ij} A_i \qquad (1 \leqslant j \leqslant n)$$

若 $i \notin I(x)$，我们取 $D_{ij} = 0$. 因而

$$A_j = \sum_{i=1}^{n} D_{ij} A_i$$

或

$$A = AD$$

我们定义

$$d = D^{\mathrm{T}} c$$

其中 c 是出现在目标函数中的向量. 注意 D 和 d 与 x 有关并且将在算法的进程中改变.

对任意的指标 $q \notin I(x)$ 及任意的 $\lambda \in \mathbb{R}$,

$$\begin{aligned} b = Ax &= \sum_{i=1}^{n} x_i A_i + \lambda A_q - \lambda \sum_{i=1}^{n} D_{iq} A_i \\ &= \sum_{i=1}^{n} (x_i - \lambda D_{iq} + \lambda \delta_{iq}) A_i \\ &\equiv \sum_{i=1}^{n} y_i A_i \end{aligned}$$

因此，我们有

$$Ay = b$$

其中 $y = x - \lambda D_q + \lambda e_q$，这里 D_q 表示 D 的第 q 列. 现在我们的目标是选择 q 和 λ 使 y 为 K 的极值点(即 $y \geqslant 0$)且 $c^{\mathrm{T}} y > c^{\mathrm{T}} x$.

因为 $q \notin I(x)$，$D_{qq} = 0$ 且 $x_q = 0$，因此，$y_q = \lambda$. 因为 y 是一个可行点，所以我们需要 $\lambda \geqslant 0$,现在计算

$$\begin{aligned} c^{\mathrm{T}} y &= \sum_{i=1}^{n} c_i x_i - \lambda \sum_{i=1}^{n} c_i D_{iq} + \lambda \sum_{i=1}^{n} c_i \delta_{iq} \\ &= c^{\mathrm{T}} x - \lambda c^{\mathrm{T}} D_q + \lambda c_q \\ &= c^{\mathrm{T}} x + \lambda (c_q - e_q) \end{aligned}$$

为递增目标函数的值，我们将选取 q 使得 $c_q > e_q$. 若 $c \leqslant e$，则 q 不存在，计算中止；x 是一个解. 否则，通常选取 q 使得 $c_q - e_q$ 尽可能大，从此 q 固定. [703]

选取 λ 使 $\lambda > 0$ 且使 $I(y)$ 至多有 m 个元素. 从 y 的定义，我们看到 $I(y) \subseteqq I(x) \cup \{q\}$. 因为 $I(x)$ 恰有 m 个元素，所以选取 λ 使 $x_i - \lambda D_{iq}$ 项之一为 0，显然其他项 $\geqslant 0$. 然而，首先看到若 $1 \leqslant i \leqslant n$ 时 $D_{iq} \leqslant 0$，则对一切 $\lambda > 0$，$y \in K$. 由前面的 $c^{\mathrm{T}} y$ 的公式，我们看到此时 $\lim\limits_{\lambda \to \infty} c^{\mathrm{T}} y = +\infty$. 因此，由于目标函数在 K 上无界，所以解不存在. 若对一个或多个 i 的值有 $D_{iq} > 0$，则我们认为 λ 从值 0 递增. 一开始，当 $i \in I(x)$ 时，$x_i - \lambda D_{iq} > 0$. 这些项的某些项是递减的，并且当它们中的某一项变为 0 时，我们取相应的 λ. 形式上，

$$\lambda = \min\left\{\frac{x_i}{D_{iq}} : x_i > 0, D_{iq} > 0\right\}$$

因此完全地确定了 y，我们要验证它就是一个极值点. 首先，设 $\lambda = x_p / D_{pq}$，其中 p 是使 $x_p > 0$ 和 $D_{pq} > 0$ 的指标. 因而

$$I(y) \subseteqq I(x) \cup \{q\} \setminus \{p\}$$

由定理 1，证明集合

$$\{A_i : i \in I(x) \bigcup \{q\} \setminus \{p\}\}$$

线性无关就足够了．假设

$$\sum_{i \in I(x)} \beta_i A_i + \beta_q A_q = 0, \quad \beta_p = 0$$

若 $\beta_q = 0$，则上式化为

$$\sum_{i \in I(x)} \beta_i A_i = 0$$

于是$\{A_i : i \in I(x)\}$的无关性推出 $\beta_i = 0$，$i \in I(x)$．因此，此时一切 β_i 都为 0．所以我们可以着手讨论 $\beta_q \neq 0$ 的情形．由齐性，可以假定 $\beta_q = -1$，等式现在可记为

$$A_q = \sum_{i \in I(x)} \beta_i A_i, \beta_p = 0$$

我们也有

$$A_q = \sum_{i=1}^{n} D_{iq} A_i = \sum_{i \in I(x)} D_{iq} A_i$$

由$\{A_i : i \in I(x)\}$的线性无关性，我们得到 $\beta_i = D_{iq}$．但是这不能成立，正如前面提到的，因为 $\beta_p = 0$ 而 $D_{pq} > 0$．因此，β_q 必为 0．至此，我们已证明 $y \in K$ 且它是一个极值点．

704

一个细节留待证明：若 $c \leqslant d$，则 x 是一个解．设 u 是任意可行点，则 $Ax = b = Au = A(Du)$，因为 Ax 和 ADu 是$\{A_i : i \in I(x)\}$的线性组合，由此可得 $x = Du$，于是我们有条件 $c^{\mathrm{T}} u \leqslant d^{\mathrm{T}} u = c^{\mathrm{T}} Du = c^{\mathrm{T}} x$，这就是希望证明的．

例 1 我们将对一个具体的例子说明单纯形法如何操作．考虑问题

$$\text{Max}: F(x) = x_1 + 2x_2 + x_3$$

$$\text{约束}: \begin{cases} x_1 + x_2 + x_5 = 1 \\ x_1 + x_3 + x_4 + x_5 = 1 \\ x_i \geqslant 0 \qquad (1 \leqslant i \leqslant 5) \end{cases}$$

解 已知的数据为

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 \end{bmatrix} \quad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad c^{\mathrm{T}} = [1 \quad 2 \quad 1 \quad 0 \quad 0]$$

开始用 $x = (0, 1, 0, 1, 0)^{\mathrm{T}}$ 和 $I(x) = \{2, 4\}$．注意到$\{A_2, A_4\}$是$\mathbb{R}^2$ 的基，因此，由定理 1 知，x 是可行集的一个极值点，或者说是一个基本可行点．

A 的每一列是 A_2 和 A_4 的线性组合，事实上，当然我们有 $A_1 = A_2 + A_4$，$A_2 = A_2$，$A_3 = A_4$，$A_4 = A_4$ 以及 $A_5 = A_2 + A_4$，所以 D 阵为

$$D = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

向量 d^{T} 是 D 的行 D^i 的线性组合：

$$d^{\mathrm{T}} = c^{\mathrm{T}} D = \sum_{i=1}^{n} c_i D^i = D^1 + 2D^2 + D^3 = [2 \quad 2 \quad 0 \quad 0 \quad 2]^{\mathrm{T}}$$

向量 $c-d$ 为

$$c-d=[-1\quad 0\quad 1\quad 0\quad -2]^{\mathrm{T}}$$

仅有一个分量是正的且 $q=3$，向量 $x-\lambda D_q$（D_q 表示 D 的第 q 列）为

$$\begin{aligned} x-\lambda D_3 &= [0\quad 1\quad 0\quad 1\quad 0]^{\mathrm{T}}-\lambda[0\quad 0\quad 0\quad 1\quad 0]^{\mathrm{T}} \\ &= [0\quad 1\quad 0\quad 1-\lambda\quad 0]^{\mathrm{T}} \end{aligned}$$

705

取 $\lambda=1$，因而 y 向量为

$$y=x-\lambda D_3+\lambda e_3=[0\quad 1\quad 1\quad 0\quad 0]^{\mathrm{T}}$$

现在用 y 代替 x 重复这个过程. 这里没有给出细节，我们在下一步求出 $d=(3,\ 2,\ 1,\ 1,\ 3)^{\mathrm{T}}$，因为 $c\leqslant d$，所以 y 是一个解且 $F(y)=3$. ■

10.4.3 表格法

单纯形法的实际做法通常是在一个**表格**中列出数据，然后按照一定的规则依次修改它们来完成的. 我们将用一个适当大小的例子来说明这个算法的结构：

$$\mathrm{Max}: F(x)=6x_1+14x_2$$

$$\text{约束}:\begin{cases} 2x_1+x_2\leqslant 12 \\ 2x_1+3x_2\leqslant 15 \\ x_1+7x_2\leqslant 21 \\ x_1\geqslant 0, x_2\geqslant 0 \end{cases}$$

在单纯形法的准备阶段，我们引入松弛变量并改写问题如下

$$\mathrm{Max}: F(x)=6x_1+14x_2+0x_3+0x_4+0x_5$$

$$\text{约束}:\begin{cases} 2x_1+x_2+x_3=12 \\ 2x_1+3x_2+x_4=15 \\ x_1+7x_2+x_5=21 \\ x_1\geqslant 0, x_2\geqslant 0, x_3\geqslant 0, x_4\geqslant 0, x_5\geqslant 0 \end{cases}$$

因此，我们有

$$\mathrm{Max}: F(x)=(6,14,0,0,0)x$$

$$\text{约束}:\begin{cases} \begin{bmatrix} 2 & 1 & 1 & 0 & 0 \\ 2 & 3 & 0 & 1 & 0 \\ 1 & 7 & 0 & 0 & 1 \end{bmatrix} x=\begin{bmatrix} 12 \\ 15 \\ 21 \end{bmatrix} \\ x=(x_1,x_2,x_3,x_4,x_5)^{\mathrm{T}}\geqslant 0 \end{cases}$$

第一个向量是 $x=(0,\ 0,\ 12,\ 15,\ 21)^{\mathrm{T}}$. 把所有这些数据集中在第一个表格中

6	14	0	0	0	0
2	1	1	0	0	12
2	3	0	1	0	15
1	7	0	0	1	21
0	0	12	15	21	

706

单纯形法的每一步从一个表格开始，顶部的行包含目标函数 F 的系数，$F(x)=c^{\mathrm{T}}x$ 的当前值显示在右上角. 表格中的下面 m 行表示体现等式约束的方程组，记住对这个方程组执行初等

行运算不改变解集. 表格的最后一行包含当前的 x 向量. 注意利用顶部和底部行容易计算 $F(x)=c^{\mathrm{T}}x$. 前述的表格具有一般形式

c^{T}	0	$F(x)$
A	I	b
x(非基本的)	x(基本的)	

10.4.4 表格法则

出现在单纯形法中的表格必须满足下列五条法则：

1. 向量 x 必须满足等式约束 $Ax=b$.

2. 向量 x 必须满足不等式 $x\geqslant 0$.

3. x 的 n 个分量(指它的**非基本变量**)为 0，其余的 m 个分量通常非零，它们是指定的**基本变量**.（这里 n 和 m 对应于松弛变量引进之前原问题有关的值.）

4. 在定义约束的矩阵中，每个基本变量仅出现在一行中.

5. 目标函数 F 必须只用非基本变量表示.

10.4.5 进一步说明

在上例第一个表格中，基本变量是 x_3，x_4 和 x_5，非基本变量是 x_1 和 x_2，我们立刻看到 5 条法则对此表格全部成立.

在每一步，考察当前的表格看看是否允许一个非基本变量变为一个基本变量来增加 $F(x)$ 的值. 在我们的例子中看到，如果允许 x_1 或 x_2 增加(通过调整 x_3，x_4，x_5 来补偿)，则 $F(x)$ 的值将确实增加. 因为 F 中的系数 14 比系数 6 大，所以在 x_2 中增加一个单位比在 x_1 中增加一个单位可以更快地使 $F(x)$增加. 因此，我们控制 x_1 使之固定为 0，允许 x_2 尽可能多的增加. 这些约束适合：

$$0\leqslant x_3=12-x_2$$
$$0\leqslant x_4=15-3x_2$$
$$0\leqslant x_5=21-7x_2$$

707

这些约束告诉我们

$$x_2\leqslant 12 \qquad x_2\leqslant 5 \qquad x_2\leqslant 3$$

这些约束中最必须遵守的是不等式 $x_2\leqslant 3$，所以允许 x_2 增加到 3. 通过 3 个给定的约束得到 x_3，x_4 和 x_5 产生的值. 因此，我们的新 x 向量是

$$x=[0\quad 3\quad 9\quad 6\quad 0]^{\mathrm{T}}$$

新的基本变量是 x_2，x_3 和 x_4，现在必须确定新表格与前述 5 条法则一致. 为满足法则 5，我们注意到 $x_2=(21-x_5)/7$. 当把它代入到 F 中后，可求出目标函数的新形式：

$$\begin{aligned}F(x)&=6x_1+14x_2\\&=6x_1+14(21-x_5)/7=6x_1-2x_5+42\end{aligned}$$

为满足法则 4，使用 7 作为主元素应用高斯消元法(初等行运算). 其目的是除了一个方程外，把 x_2 从所有其他方程中消去，所有这些工作实施后，第 1 步便完成. 用第二个表格开始第 2 步，该表格是

$$
\begin{array}{ccccc|c}
6 & 0 & 0 & 0 & -2 & 42 \\
\hline
\frac{13}{7} & 0 & 1 & 0 & -\frac{1}{7} & 9 \\
\frac{11}{7} & 0 & 0 & 1 & -\frac{3}{7} & 6 \\
1 & 7 & 0 & 0 & 1 & 21 \\
\hline
0 & 3 & 9 & 6 & 0 &
\end{array}
$$

现在呈现的情况类似于开始的情况．非基本变量是 x_1 和 x_5，在 x_5 中的任何增加将减少 $F(x)$，所以现在允许变为基本变量的是 x_1．因此，我们控制 x_5 使之固定为 0，并且允许 x_1 尽可能多的增加．这些约束适合：

$$0 \leqslant x_3 = 9-(13/7)x_1$$

$$0 \leqslant x_4 = 6-(11/7)x_1$$

$$0 \leqslant 7x_2 = 21 - x_1$$

这些不等式导致

$$x_1 \leqslant 63/13 \qquad x_1 \leqslant 42/11 \qquad x_1 \leqslant 21$$

新的基本变量 x_1 只允许增加到 42/11，从新的表格中或直接从上面的约束方程中计算出新的 x_3，x_4 和 x_2 的值，新 x 向量是

$$x = [42/11 \quad 27/11 \quad 21/2 \quad 0 \quad 0]^{\mathrm{T}}$$

现在新的非基本变量是 x_4 和 x_5，为满足法则 5，我们使用替换 $x_1=(7/11)(6-x_4)$，于是 [708]

$$
\begin{aligned}
F(x) &= 6x_1 - 2x_5 + 42 \\
&= (42/11)(6-x_4) - 2x_5 + 42 \\
&= -(42/11)x_4 - 2x_5 + 714/11
\end{aligned}
$$

因为 F 中的两个系数都是负的，所以不必计算第三个表格．因为非基本变量 x_4 和 x_5 中无论哪一个变为基本变量都不减少 $F(x)$，所以这标志着当前的 x 是一个解．因此，原问题的最大值是 $F(42/11,\ 27/11)=630/11$.

10.4.6 小结

在这个例题及给出的说明的基础上，我们可以对任何给定的表格所做的工作总结如下：

1. 若 F 中的所有系数(即表格中的顶部行)为$\leqslant 0$，则当前的 x 是解．

2. 选择 F 中其系数是正的且尽可能大的非基本变量，这个变量变成一个新的基本变量，称它为 x_j.

3. 用所在行的新的基本变量的系数 a_{ij} 除每个 b_i，这个赋值到新的基本变量的值是这些比中的最小者．因此，当 b_k/a_{kj} 最小时，我们取 $x_j=b_k/a_{kj}$.

4. 用主元素 a_{kj}，应用高斯消元法在 A 的第 j 列中产生 0.

10.4.7 工作量估计

在单纯形法的实际应用中，工作量的理论界和实际工作量之间有相当大的差别．单纯形步数的上界是二项式系数

$$\binom{n}{m}$$

然而实际上步数通常不超过 $2m$. 即使在一个适当大小的问题中，譬如说 $n=300$，$m=100$，前面的二项式系数的值是天文数字. 事实上，利用斯特林公式(见习题 10.4.1)，我们有

$$\binom{300}{100}=\frac{300!}{200!100!}\approx 4\times 10^{81}$$

10.4.8 其他算法

线性规划的一个新算法是由 Karmarkar[1984]发表的，新算法声称，当变量个数达到 709 15 000或更多时，它比单纯形法优越.

卡马卡算法的工作量是问题大小的一个多项式函数，而单纯形法的工作量是问题大小的一个指数函数. 这本身并不意味着新算法是较好的，但卡马卡算法多年来被证明对许多极大型问题是优越的. 现在认为对大型线性规划问题，它和其他的内点法与单纯形法不相上下. 先前由 Khachian 给出的算法也有问题大小的多项式工作量函数，但是那个方法绝不比单纯形法有竞争力，因为它在逐次的步骤中需要越来越高的精度.

习题 10.4

1. 大家应该知道斯特林公式，它给出 $n!$ 的估计：

$$n!\approx\sqrt{2\pi n}\left[\frac{n}{e}\right]^n$$

利用它导出近似公式

$$\binom{n}{m}=\frac{n!}{m!(n-m)!}\approx\sqrt{\frac{n}{2\pi m(n-m)}}\left[\frac{n}{n-m}\right]^n\left[\frac{n-m}{m}\right]^m$$

2. (续)利用上题验证

$$\binom{300}{100}\approx 4\times 10^{81}$$

3. 利用本节中提出的单纯形法的要点和表格求解下列问题：

$$\text{Max}: F(x)=2x_1-3x_2$$

$$\text{约束}:\begin{cases}2x_1+5x_2\geqslant 10\\ x_1+8x_2\leqslant 24\\ x_1\geqslant 0, x_2\geqslant 0\end{cases}$$

4. 仔细考虑本节中给出的单纯形法，证明：若 x 是一个解，则 $c\leqslant d$.

5. a. 利用表格重复求解本节的第一个例题.

   b. 利用单纯形法重复求解第二个例题.

6. 利用表格法，求解下列问题：

$$\text{Max}: F(x)=6x_1+14x_2$$

$$\text{约束}:\begin{cases}x_1+x_2\leqslant 12\\ 2x_1+3x_2\leqslant 15\\ x_1+7x_2\leqslant 21\\ x_1\geqslant 0, x_2\geqslant 0\end{cases}$$

710 改变第二个不等式为 $2x_1+3x_2\geqslant 15$，再重复求解.

第11章 最 优 化

11.0 概述

我们把**最优化**理解为在特定的区域上求某个多实变量实值函数的极小. 当我们说*求极小*时，通常理解为求函数的最小值以及最小值出现的点. 当然，上一章专门讨论了一种重要的极小化类型. 即函数是线性的，并用一组线性不等式来描述定义域. 因为这些情况的特殊结构，所以使用特定的方法. 它们允许处理上千个变量和上千个不等式约束.

当我们脱离线性函数这种宽松环境时，则比较混乱. 即使求一个没有任何约束的单变量的函数的极小，也可能具有挑战性. 当问题中的函数有许多纯粹的**局部极小**时，会出现极大的困难，因为我们真正需要的是**整体极小点**. 在处理有若干约束的多变量函数的问题中可能会出现更多的困难. Nocedal-Wright[1999]是一本有关整个最优化领域的教科书，几本其他一般性的著作在本章的后面列出.

最优化学科自然地分成许多分支，它与所研究问题的结构有关. 我们已经提到的线性规划是一个重要的分支，另一个分支是凸分析，其中无论是函数还是约束都呈现凸性，Borwein-Lewis[2000]是专门讨论这个主题的一本新书. 一个特别的主题是其最优点必须在预先指定的格上(如在$\mathbb{R}^n$中的整向量的格)，它被研究了40年或更长时间. 对这种问题所用方法的最新资料是Cornuejols[2000]. 若干本书讨论了实施最优化算法的计算机程序，如Bhatti[2000].

711

互联网是关于最优化的资讯和计算机程序的一个宽广的来源. 例如，站点

http://plato.la.asu.edu/guide.html

提供选择适当的最优化软件的决策树. 专门讨论遗传算法的站点是

http://www.aic.nrl.navy.mil:80/galist/

也可见本书附录A. 新的讨论可在

http://www.netlib.org/na-digest/

中的NA-Digest中找到. 因为网站时常处于流动的状态，所以用户可以依靠搜索引擎查找帮助的原始资料，可以提出许多关键词，诸如optimization，simulated annealing，genetic algorithms等等.

在本章中使用的是标准的术语，我们从一个预先指定的n个实变量的实值函数f入手，求f的*整体极小点*. 这意味着对一切点x，点p满足$f(p)\leqslant f(x)$. 相反地，**局部极小**是一个点q使得对某个包含q的开集和在上述开集中的一切x有$f(q)\leqslant f(x)$. 这是一个较弱的概念，因为整体极小必定是一个局部极小. 求局部极小是较容易的，而要弄明白已知的(任何)局部极小点是整体极小点通常不是一件简单的事情.

$\mathbb{R}^n$中的一个向量通常简记为x，如果需要分量，我们用$x=(x_1, x_2, \cdots, x_n)^T$. 如果讨论点(向量)序列，则用$x^{(1)}, x^{(2)}, \cdots$，等等表示序列.

因为多变量算法通常需要线搜索，所以单变量的情形显然是一个初始点，这就是下一节的

主题.

11.1 单变量情况

一个向量空间的一条线用单个实变量来刻画，它具有形式

$$\{u+tv : t \in \mathbb{R}\}$$

其中 u 和 v 是环绕空间中的特定向量($v\neq 0$). 若 F 是定义在向量空间上的实值函数，则由等式 $f(t)=F(u+tv)$ 定义的函数 f 是一个实变量函数. 如果线到达靠近 F 的极小点，则可以用极小化较简单的函数 f 开始我们的搜索. 因此，我们将考虑求任意函数 $f:\mathbb{R}\to\mathbb{R}$ 极小点的一般问题.

712

若函数 f 处处可微，则在一个局部极小点 q 上，必有 $f'(q)=0$. 在微积分中学过这个内容，处理这个问题的方法是确定 f 的导数为零的所有点. 检验每个这样的点，看看它是否为**局部极小点**、**局部极大点**或**鞍点**. 此时，鞍点定义为一个点，在这个点的函数值上升或下降与选择从这个点离开的方向有关.

这些考虑提出了一个重要的问题. 我们需要两类算法，第一类使用 f 的导数(如果它有导数)，第二类不使用 f 的导数(即使它有导数). 研究者通常在选择代码时需要机动性. 当处理多变量函数时，这个区别更为重要，因为 n 个变量的函数有 n 个偏导数，每个偏导数在局部极小点上都必须等于 0. 在一个程序中使用的这些导数的编程可能花费太多的人力，并且当算法在计算机上运行时，导数的计算可能很费机时.

如果由于这样或那样的原因，不可利用 f 的导数，则可采用一个简单的搜索过程. 例如，可选取一个步长 h(大致地根据 f 怎样变化的某些知识)，然后可以计算值 $f(kh)$，$k=0, \pm 1, \pm 2, \cdots$. 用这个方法用户得到 $f(x)$ 怎样变化的一些资料，以及极小点可能隐藏在哪里. 对任意特定的 h 值，容易构造一个函数使得一个基于这种搜索的算法不能工作. 但是如果函数不服从它可能怎样变化的限制的话，这几乎对任何算法都成立. 只需要连续性是太弱了. 更有用的是导数的界，例如对一切 x，$|f'(x)|\leqslant M$. 如果 $f(x)$ 在两个点 a 和 b 上的数值已知，其中 $a<b$，并且 M 是已知的，则在区间 $[a, b]$ 上，

$$f(x) \geqslant \min\{f(a), f(b)\} - \frac{1}{2}(b-a)M \tag{1}$$

用中值定理来帮助证明此式. 例如，当 x 在区间 $[a, b]$ 的左半部分时，

$$f(x)-f(a) = f'(\xi)(x-a) \geqslant -M(x-a) \geqslant -\frac{1}{2}(b-a)M$$

由此立即得到不等式(1). 当 x 在区间的右半部分中的证明是类似的. 如果这样的函数 f 已在点 $x^{(k)}=kh$ 上采样，则前面的分析指出

$$\min_k f(x^{(k)}) \geqslant \inf_x f(x) \geqslant \min_k f(x^{(k)}) - \frac{1}{2}hM$$

现在搜索可以缩小到区间 $[x^{(j)}, x^{(j+1)}]$，使得

$$\min\{f(x^{(j)}), f(x^{(j+1)})\} \leqslant \min_k f(x^{(k)}) + \frac{1}{2}hM$$

当然，如果区间$[x^{(j)}, x^{(j+1)}]$违反这个不等式，则

$$\inf_{x^{(j)}\leqslant x\leqslant x^{(j+1)}} f(x) \geqslant \min\{f(x^{(j)}), f(x^{(j+1)})\} - \frac{1}{2}hM > \min_k f(x^{(k)})$$

713

因为$f(x)$在该区间上的极小值大于f在一个采样点上的值，所以可以忽略不计这个区间.

若不知道函数导数的性质，则可以采用一个使用函数的较弱性质的算法. 即可以假定f是连续的并且在求极小值的区间$[a, b]$上是单峰的. **单峰**意味f在区间$[a, b]$上有一个单独的局部极小值. 如果f是一个这样的函数，并且，若x^*是f在$[a, b]$上的一个极小点，则f在$[a, x^*]$中是递减的而在$[x^*, b]$中是递增的. 为证明这点，相反地假定f在$[x^*, b]$中不递增. 于是，存在点u和v，使得$x^* \leqslant u < v \leqslant b$且$f(u) \geqslant f(v)$. 选择$x^{**}$为$f$在$[u, b]$中的极小点，如果$u$是$x^{**}$的一个可能的选择，则选择$v$为$x^{**}$来代替$u$. 因为$f(v) \leqslant f(u)$，所以这是允许的. 现在$x^{**}$显然是$f$的另一个局部极小点，因为它是$f$在$[u, b]$上的极小点而且它不是$u$. 另一种情形的证明是类似的.

一个称为**黄金分割搜索**的算法可应用于连续单峰函数. 它利用数

$$r = \frac{1}{2}(\sqrt{5} - 1) = 0.618\,033\,988\,7\cdots$$

这个数被古希腊人看成是绘画或建筑方面美学的最优比率. 它是方程$r^2 = 1 - r$的一个根.

在这个算法的每一步，从前面的计算可知区间$[a, b]$是可用的. 取$x = a + r(b-a)$和$y = a + r^2(b-a)$，需要函数值$u = f(x)$和$v = f(y)$，但是我们将看到，在算法开始启动以后，对每个新区间，只需计算一个新的f值.

若$u > v$，则由f的单峰性，f的最小值必在$[a, x]$中. 这个区间变成下一步的输入区间. 注意现在已经有这个区间的一个内点上的f值，即$v = f(y)$. 容易验证，在这个区间内y的位置为$y = a + r(x-a)$. 在下一步中，y扮演老的x的角色，并且必须计算在$a + r^2(x-a)$的f值. 此时，这些替换应该按严格的顺序执行：

$$\begin{aligned} &b \leftarrow x \\ &x \leftarrow y \\ &u \leftarrow v \\ &y \leftarrow a + r^2(b-a) \\ &v \leftarrow f(y) \end{aligned}$$

另一种情形，$u \leqslant v$且最小点必在区间$[y, b]$中. 我们注意到$x = y + r^2(b-y)$，因而所需的替换按严格的顺序是

$$\begin{aligned} &a \leftarrow y \\ &y \leftarrow x \\ &v \leftarrow u \\ &x \leftarrow a + r(b-a) \\ &u \leftarrow f(x) \end{aligned}$$

714

这个算法的做法使人联想起求函数零点的对分法，该方法中，逐次的区间长度以因子0.5递减，但是在黄金分割搜索中，区间的宽度以近似于0.62的因子递减，这不是令人满意的.

为说明当导数可以利用时所用算法的种类，回忆对从微积分得到的方法所作的注记：仅仅求一个使 $f'(x)=0$ 的点. 求出这样的一个点后，确定它是否为一个局部极大的，一个局部极小点，或者一个鞍点. 为求 f' 的零点，我们可以借助于牛顿法，这导致迭代

$$x^{(k+1)} = x^{(k)} - \frac{f'(x^{(k)})}{f''(x^{(k)})} \tag{2}$$

可以证明(见习题 11.1.4)(2)式中的算法完全相当于用埃尔米特插值构造的二次函数 Q 的局部拟合函数 f. 即要求 Q 满足

$$Q(x^{(k)}) = f(x^{(k)}) \quad Q'(x^{(k)}) = f'(x^{(k)}) \quad Q''(x^{(k)}) = f''(x^{(k)}) \tag{3}$$

完成这些要求后，取 Q 的极小点为 $x^{(k+1)}$. 实际上取使 $Q'(x^{(k+1)})=0$ 的点，这样有一个惊奇的可能性等待着用户，Q 可能在 $x^{(k+1)}$ 有一个局部极大值!

这种策略的变种是利用割线法(代替牛顿法)去求 f' 零点附近的点. 这就避免了求 f'' 的公式及对它编写程序的必要性. 当然，其他的方法如对分法可以用于求出 f' 的零点.

已经提出用一个插值多项式近似函数的可能性，我们可以考虑通常的逐点插值而非埃尔米特插值. 此时，在三个最新的点 $x^{(k)}$，$x^{(k-1)}$ 和 $x^{(k-2)}$ 上用一个二次多项式插值 f. 或许所得的二次函数 Q 局部地酷似 f 的特性，并且 Q 的极小点可以取作下一点 $x^{(k+1)}$. 以一个稳健的方式实施这个算法，我们必须进一步决定如果 $x^{(k+1)}$ 不是前面的点的改进时要做什么. 关于这个课题，可以参见 Powell[1964]，本算法在 Walsh[1975]中也作了讨论，也可参见 Dahlquist-Björck[1974]. 基于插值的其他算法是可用的，诸如由 Davidon 给出的一个在每一步使用三次多项式的算法，这个算法在 Walsh[1975]和 Buchanan-Turner[1992]中作了讨论.

习题 11.1

1. 证明在黄金分割搜索算法中，逐次区间的长度以近似于 0.62 的因子递减.
2. 考虑极小化 f 的算法，从用对分法求 f' 的根着手，说明这个简单的算法比黄金分割搜索快多少?
3. 假定已知 $|f'(x)|$ 在区间$[a, b]$上的一个上界 M，描述一个该区间上计算函数 f 的极小点的算法. 用户指定使用的相等的子区间数 N，采样点是 $x^{(i)}=a+ih$，其中 $h=(b-a)/N$，$0\leqslant i\leqslant N$. 根据本节的分析，在计算结束时，代码报告可能是极小点的一切子区间. 如果有兴趣，考虑迭代过程，即在第二阶段处理第一阶段残存的区间，等等.

715

4. 证明(2)式的算法等同于像(3)式中由埃尔米特插值导出的算法.

11.2 下降法

前面已经提出几个极小化线性函数的算法，我们继续考虑极小化 n 个实变量的函数这种更具挑战性的问题. 需要一些微积分的术语. 若 f 是一个$\mathbb{R}^n$ 到$\mathbb{R}$ 的函数，则 f 在点 x 上的**梯度**是向量 G，它的分量是

$$G_i = G_i(x) = \frac{\partial f(x)}{\partial x_i} \qquad (1 \leqslant i \leqslant n)$$

它也用$\nabla f(x)$表示，或简单地用 $f'(x)$表示，后面的表达式是**弗雷歇导数**的标准记号. 它解释为一个线性映射，它在 u 上的值是 $u^{\mathrm{T}} f'(x)$.

f 在 x 的**黑塞矩阵**是矩阵 H，它的元素是

$$H_{ij}=H_{ij}(x)=\frac{\partial^2 f(x)}{\partial x_i \partial x_j} \qquad (1\leqslant i,j\leqslant n)$$

它也可解释为一个弗雷歇导数，所以它可用 $f''(x)$ 表示. 用梯度和黑塞矩阵，我们可以写出 f 的**泰勒展开式**中的一次项和二次项：

$$f(x+h)=f(x)+G(x)^{\mathrm{T}}h+\frac{1}{2}h^{\mathrm{T}}H(x)h+\cdots \tag{1}$$

梯度向量点在**最速上升**方向，并且它的相反的点在**最速下降**方向. 这个重要的事实通过考虑任意的单位向量 u，并考虑函数在 x 沿 u 方向如何局部变化来证明. 作为方向导数这是容易计算的；它是

$$\frac{\mathrm{d}}{\mathrm{d}t}f(x+tu)\bigg|_{t=0}$$

利用(1)式中的泰勒公式，h 用 tu 代替，我们得到

$$f(x+tu)=f(x)+tG(x)^{\mathrm{T}}u+\frac{1}{2}t^2u^{\mathrm{T}}H(x)u+\cdots$$

由此

$$\frac{\mathrm{d}}{\mathrm{d}t}f(x+tu)=G(x)^{\mathrm{T}}u+tu^{\mathrm{T}}H(x)u+\cdots$$

716

另一方面，可以利用复合函数微分的链式法则写成

$$\frac{\mathrm{d}}{\mathrm{d}t}f(x+tu)=u^{\mathrm{T}}f'(x+tu)=u^{\mathrm{T}}G(x+tu)=u^{\mathrm{T}}\nabla f(x+tu) \tag{2}$$

这就避免了求助于泰勒公式. 取 $t=0$，我们得到 $G(x)^{\mathrm{T}}u$ 作为 f 在 x 沿方向 u 的**变化率**. 由柯西-施瓦茨不等式，这个变化率不超过 $\|G(x)\|\ \|u\|=\|G(x)\|$. 另一方面，取 u 是沿 $G(x)$ 方向的单位向量，我们可以达到这个上界.

求 f 极小点的一个显而易见的策略是在任意点 x 开始，确定 f 以最快的速率局部递减的方向，即，$-G(x)$ 的方向. 可以在射线 $\{x-tG(x):t>0\}$ 上执行线搜索. 在这条射线上 f 的极小点上，我们再开始计算新的最速下降方向，等等.

这个称为**最速下降**的算法尽管在理论上有吸引力，但是它常常执行得很差，因为它花费许多时间走锯齿形方向而不转向整体极小点. 甚至对二次函数也可看到这种情况，见 4.7 节图 4-3. 容易证明这个方法产生的逐次的方向必须互相正交，所以锯齿形是这个方法的一个不可避免的性质. 较好的是考察一条线，它是最近实施的好几个方向的平均或者从开始重新考虑如何选择适当的线搜索的方向. 这一主题在后面几节中继续讨论.

在最速下降中逐次方向的正交性包含在下列引理中.

引理 1(最速下降中的正交性引理) *若函数 f 在线 $\{w+tu:t\in\mathbb{R}\}$ 上的极小值在 x 出现，则 u 垂直于 f 在 x 的梯度.*

证明 点 x 在所描述的线上，因此点 $x+tu$ 也在那条线上. 定义 $g(t)=f(x+tu)$，则 g 的极小值出现在 0 处，并且 $g'(0)=0$. 由(2)式，$g'(t)=u^{\mathrm{T}}G(x+tu)$，由此，$0=g'(0)=u^{\mathrm{T}}G(x)$. ■

创造几个试验函数为最优化程序提供**基准**，这些函数通常显示变量的数量级范围或者导致

极小点的长而扭曲的凹谷．这里是 Scales[1985]给出的五个这样的函数，每种情况给出一个麻烦的初始点．

Scales 函数：

$$F(x) = e^x + 0.01/x$$

717 初始点为 1.0.

罗森布罗克函数(见图 11-1)：

$$F(x) = 100(x_1^2 - x_2)^2 + (1 - x_1)^2$$

初始点为(−1.2，1.0)．

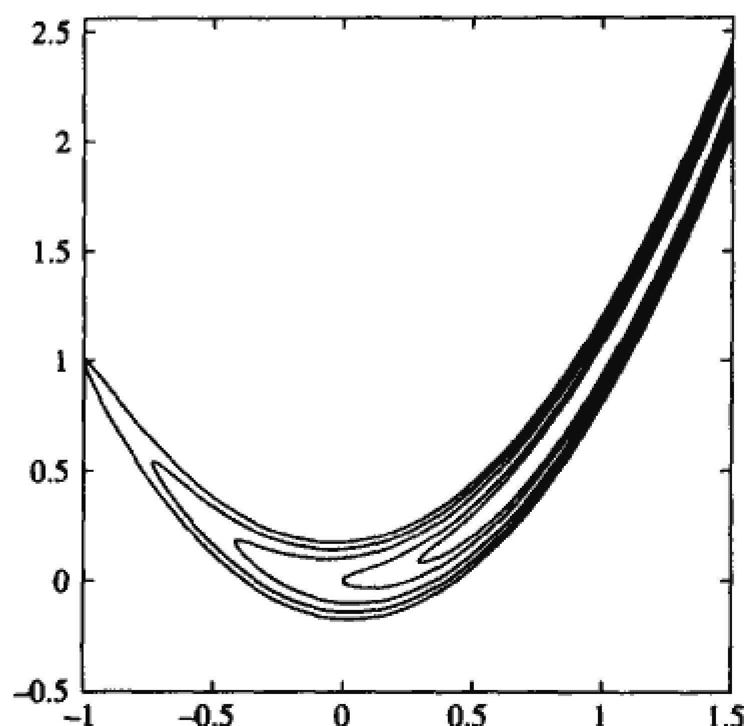


图 11-1 罗森布罗克函数的围道线

Fletcher 和 Powell 的**螺旋线函数：**

$$100[(x_3 - 10\theta)^2 + (r - 1)^2] + x_3^2$$

初始点为(−1，0，0)．这里 $r=\sqrt{x_1^2+x_2^2}$，当 $x_1 \geqslant 0$ 时 $2\pi\theta = \tan^{-1}(x_2/x_1)$，当 $x_1 < 0$ 时 $2\pi\theta = \pi + \tan^{-1}(x_2/x_1)$．

Wood 函数：

$$100(x_2 - x_1^2)^2 + (1 - x_1)^2 + 90(x_4 - x_3^2)^2 +$$
$$(1 - x_3)^2 + 10.1[(x_2 - x_1)^2 + (x_4 - 1)^2] + 19.8(x_2 - 1)(x_4 - 1)$$

初始点为(−3，−1，−3，−1)．

鲍威尔奇异函数：

$$F(x) = (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^4 + 10(x_1 - x_4)^4$$

初始点为(3，−1，0，1)．

习题 11.2

1. 求课本中列出的试验函数的梯度和黑塞矩阵．

718

2. 证明：F 的黑塞矩阵是 ∇f 的雅可比行列式．

11.3 二次目标函数的分析

极小化问题中常用的策略是假定我们的函数近似地是二次多项式. n 个变量的一般二次函数具有形式

$$f(x)=a-b^{\mathrm{T}}x+\frac{1}{2}x^{\mathrm{T}}Ax \tag{1}$$

这里，a 是一个标量，b 是一个 n 个分量的常向量，A 是 $n\times n$ 对称常数矩阵. 等式的右边(从左到右)展开成 0 次项、精确的一次项和精确的二次项. 在两个变量的情形中，后面的项具有三种类型 x_1^2，x_2^2 和 x_1x_2. 附加在 b 前面的负号使得我们的分析与 4.7 节讨论线性代数中的斜量法一致. 如果函数 f 具有极小点(不是极大点或鞍点)，则矩阵 A 必定是非负定的. 这个结论出自多变量函数标准的二阶导数检验.

为知道这个函数 f 需要精确地知道一切可利用的$\binom{n+2}{2}$或$(n+1)(n+2)/2$ 个参数. 这是 n 个变量的二次多项式空间的维数.

因为二次函数是任何给定的二次连续可微函数的一个良好的局部化模型，所以将对它们导出某些基本事实，给出大多数的细节. 首先 f 的完整形式是

$$f(x)=a-\sum_{i=1}^{n}b_ix_i+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}x_ix_jA_{ij}$$

其次，我们计算 f 的梯度，它定义为具有分量

$$\frac{\partial f}{\partial x_k}=-b_k+\frac{1}{2}\sum_{j=1}^{n}x_jA_{kj}+\frac{1}{2}\sum_{i=1}^{n}x_iA_{ik}$$

的向量. 因为矩阵 A 对称，所以这个等式变成

$$\frac{\partial f}{\partial x_k}=-b_k+\sum_{j=1}^{n}A_{kj}x_j$$

因此，梯度的向量表达式为

$$G(x)=\nabla f(x)=Ax-b \tag{2}$$

为求 f 的临界点，即 f 的导数为零的点，取 $G(x)=0$，注意到当 $Ax=b$ 时，换言之，当 $x=A^{-1}b$ 时出现这种情况. 为证明最后的逻辑步合理，必须假定 A 是正定阵而不仅仅是半正定阵.

719

我们需要一个在任何线$\{w+th:-\infty<t<\infty\}$上求 f 的极小点的公式. 我们有

$$\begin{aligned}f(w+th)&=a-b^{\mathrm{T}}(w+th)+\frac{1}{2}(w+th)^{\mathrm{T}}A(w+th)\\&=a-b^{\mathrm{T}}w-tb^{\mathrm{T}}h+\frac{1}{2}w^{\mathrm{T}}Aw+tw^{\mathrm{T}}Ah+\frac{1}{2}t^2h^{\mathrm{T}}Ah\end{aligned}$$

因为 $w^{\mathrm{T}}Ah=h^{\mathrm{T}}Aw$，因而关于 t 的导数是

$$\frac{\mathrm{d}}{\mathrm{d}t}f(w+th)=-b^{\mathrm{T}}h+w^{\mathrm{T}}Ah+th^{\mathrm{T}}Ah$$

取这个导数为零并解出 t，得到

$$t^* = \frac{(b-Aw)^{\mathrm{T}}h}{h^{\mathrm{T}}Ah}$$

于是，在给定的线上的**极小点**是

$$w + t^* h = w + \left[\frac{(b-Aw)^{\mathrm{T}}h}{h^{\mathrm{T}}Ah}\right]h \tag{3}$$

当 h 是点 w 处的负梯度方向上的点时，利用 $h=b-Aw$ 得到一个特殊情形为

$$u = w + t^* h = w + \left[\frac{(b-Aw)^{\mathrm{T}}(b-Aw)}{(b-Aw)^{\mathrm{T}}A(b-Aw)}\right](b-Aw) \tag{4}$$

现在可用这个分析描述二次函数最速下降算法中的一个单步. 从任意给定的点 w 出发，我们得到(4)式中给出的点 u. 当然，如果已知函数是二次的，则不需要利用最速下降；代之，可以直接计算临界点 $x=A^{-1}b$. 这里假定 A 和 b 是已知的，如果 A 和 b 不是可以直接利用的，则可以利用插值理论，在一组**一般位置**上的 $\binom{n+2}{2}$ 个点上插值 f 求出二次函数. 即这组点必须导致一个非奇异的插值问题，所以不一定落在任何二次流形上. 换言之，一个非零的二次函数 Q 不一定存在从而使插值点全部满足 $Q(x)=0$. 在 $n=2$ 的情形，可取的 6 个点是 0，$e^{(1)}$，$e^{(2)}$，$2e^{(1)}$，$2e^{(2)}$ 和 $e^{(1)}+e^{(2)}$，其中 $e^{(1)}=(1,0)$，$e^{(2)}=(0,1)$. 这组点的位移和适当的比例也是满足的. f 在这样一组 6 个点上的采样可以直接得到 5 个一阶和二阶偏导数的估计.

在一般的(非二次)情况下，我们假定二阶偏导数的存在性和连续性，所以可以计算黑塞矩阵 $H(x)$ 和梯度向量 $G(x)$. 这样计算上和/或分析上的代价可能是昂贵的. 但是对这些情况，求 f 的极小点的一个显而易见的候选方法是临界点的牛顿估计：

[720]

$$x^{(k+1)} = x^{(k)} - H(x^{(k)})^{-1}G(x^{(k)})$$

然而，这个牛顿迭代直到点序列进入到解的相对较小的邻域之前似乎不太满意. Dahlquist and Björck[1974，第 442 页]中所述的 Biggs 和 Dixon 算法仅用点 $v^{(k)}=H(x^{(k)})^{-1}G(x^{(k)})$ 去指定一个搜索方向. 因而，在点 $x^{(k)}$ 上，如刚才所述的算法计算 $v^{(k)}$，且在线 $\{x^{(k)}+tv^{(k)}: t\in\mathbb{R}\}$ 上执行一个搜索. 点 $x^{(k+1)}$ 是这个线搜索的结果，并且过程是重复的. 在一个**阻尼形式**中，如果 $H(x^{(k)})$ 奇异或者如果

$$G(x^{(k)})^{\mathrm{T}}v^{(k)} < \max\{0, \| G(x^{(k)}) \|^4/[G(x^{(k)})^{\mathrm{T}}H(x^{(k)})G(x^{(k)})]\}$$

则搜索方向向量 $v^{(k)}$ 用 $u^{(k)}=G(x^{(k)})$ 代替(像最速下降中一样).

11.4 二次拟合算法

在这个类型中，方法设计为通过修正函数 f 的近似黑塞矩阵或逆黑塞矩阵以保持一步一步更新 n 个变量的二次函数的模型. 一般说来，假定函数 f 的梯度可获得解析的形式，易于在任何给定点上进行计算. 黑塞矩阵不能解析地获得. 一个显而易见的迫切要求是这样的一个方法可快速地求出二次函数的极小值.

以 Davidon、Fletcher 和 Powell 名字命名的一个著名的算法是容易描述的. 为此，我们使用如下记号：用 $G(x)$ 表示 f 在 x 的梯度，黑塞矩阵 $H(x)$ 的近似逆用 J 表示，它在每一步是可变的. 这个公式中的所有向量假定是列向量，而像 v^{T} 这样的向量是一个行向量. 特别地，

vv^{T} 是一个 $n\times n$ 的秩一矩阵，而 $y^{\mathrm{T}}v$ 一个标量，其中 v 和 y 是向量.

在执行程序之前，我们提供计算任意点 x 上的函数 $f(x)$ 和梯度向量 $G(x)$ 的计算机程序. 初始点 $x^{(1)}$ 是给定的，并且取 J 为 n 阶单位矩阵 I_n.

在算法的第 $k+1$ 步，已经有了 $x^{(k)}$ 和其他数据，我们执行下列算法：

算法 1 (Davidon-Fletcher-Powell 算法)

1. 计算 $u\leftarrow -JG(x^{(k)})$.
2. 选择 t 给出 f 在射线 $x^{(k)}+tu$ 上的近似极小值.
3. 取 $v\leftarrow tu$.
4. 取 $x^{(k+1)}\leftarrow x^{(k)}+v$.
5. 计算 $y\leftarrow G(x^{(k+1)})-G(x^{(k)})$.
6. 替代 $J\leftarrow J+[vv^{\mathrm{T}}/y^{\mathrm{T}}v]-[(Jy)(Jy)^{\mathrm{T}}/y^{\mathrm{T}}Jy]$.

这个算法是共轭梯度法，它与 4.7 节的算法相似. 该节中的定理 2 可按最优化理论的语言重新叙述. 假定 f 是二次函数

$$f(x)=a-b^{\mathrm{T}}x+\frac{1}{2}x^{\mathrm{T}}Ax$$

721

用指定的一组**搜索方向** $v^{(1)}$，$v^{(2)}$，…定义一个下降算法. 这样，用一个正式的描述，我们有

1. $x^{(1)}$ 是任意的.
2. 给定 $x^{(k)}$，定义 $x^{(k+1)}$ 为射线 $x^{(k)}+tv^{(k)}$ 上的点，f 在这个点上有极小值.

4.7 节的定理 2 称，若搜索方向 $v^{(1)}$，$v^{(2)}$，…形成一个 A 正交系，则得到函数 f 的极小值不会超过第 $n+1$ 步. **A 正交性**是当 $i\neq j$ 时 $v^{(i)\mathrm{T}}Av^{(j)}=0$. Davidon-Fletcher-Powell 方法构造一个向量的 A 正交系，向量 u 出现在上面描述的算法的第 1 步中.

与这个方法有关的原文是 Davidon[1959]以及 Fletcher and Powell[1963]. Luenberger [1973]的教科书证明了 Davidon-Fletcher-Powell 算法的基本定理.

习题 11.4

证明：如果上述算法第 1 步和第 2 步中，用 $v^{(k)}$ 正交 $G(x^{(j-1)})-G(x^{(j)})$，$1\leqslant j<k$，则向量 $v^{(k)}$ 形成一个 A 正交集.

11.5 Nelder-Mead 算法

一个称为 **Nelder-Mead 算法**的方法可用于极小化函数 $f:\mathbb{R}^n\rightarrow\mathbb{R}$. 它是一个**直接搜索方法**，并且它的进行不含有函数 f 的导数和任何线搜索.

在开始计算之前，用户指定三个参数 α，β 和 γ 的值，默认值分别是 1，$\frac{1}{2}$和 1.

在这个算法的每一步，给出$\mathbb{R}^n$ 中的一组 $n+1$ 个点：

$$\{x^{(0)},x^{(1)},\cdots,x^{(n)}\}$$

这组点集应该位于$\mathbb{R}^n$ 中的**一般位置**. 这意味着 n 个点 $x^{(i)}-x^{(0)}$，$1\leqslant i\leqslant n$ 的集合是线性无关的. 由此假设可得原来点集$\{x^{(0)}$，$x^{(1)}$，…，$x^{(n)}\}$的**凸包**是一个 **n 单纯形**. 例如，2 单纯形是

$\mathbb{R}^2$ 中的一个三角形，而 3 单纯形是$\mathbb{R}^3$ 中的一个四面体．为使算法的描述尽可能简单，假定这些点被重新编号(如果必要的话)，使得

$$f(x^{(0)}) \geqslant f(x^{(1)}) \geqslant \cdots \geqslant f(x^{(n)})$$

因为我们打算极小化函数 f，所以点 $x^{(0)}$ 是当时的点集中最差的一个点(因为它产生 f 的最大值).

计算点

[722]

$$u = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$$

这是当时的单纯形相对于最差的顶点 $x^{(0)}$ 的面的**形心**．其次，计算**反射点** $v=(1+\alpha)u-\alpha x^{(0)}$.

若 $f(v)$小于 $f(x^{(n)})$，则这是一个顺利的情况，并且试探用 v 代替 $x^{(0)}$ 重新开始．然而，我们首先计算一个**扩大的反射点** $w=(1+\gamma)v-\gamma u$ 并且试验观察 $f(w)$是否小于 $f(x^{(n)})$．如果成立的话，则用 w 代替 $x^{(0)}$ 再开始．否则，用 v 代替 $x^{(0)}$，正如原来建议的那样，从一个新单纯形再开始.

假定现在 $f(v)$不小于 $f(x^{(n)})$．若 $f(v)\leqslant f(x^{(1)})$，则用 v 替代 $x^{(0)}$重新开始.

处理了所有的 $f(v)\leqslant f(x^{(1)})$的情形后，现在进一步考虑两种情况．首先，若 $f(v)\leqslant f(x^{(0)})$,则设 $b=f(x^{(0)})$且用 v 代替 $x^{(0)}$，无论是否 $f(v)\leqslant f(x^{(0)})$，计算 $w=\beta x^{(0)}+(1-\beta)u$，并试验是否 $f(w)\leqslant b$．如果成立，则用 w 代替 $x^{(0)}$ 再开始．如果不成立，则用$\frac{1}{2}(x^{(i)}+x^{(n)})$代替 $x^{(i)}$，$0\leqslant i\leqslant n-1$，缩小单纯形再开始.

下式定义的值

$$\frac{f(x^{(0)})-f(x^{(n)})}{|f(x^{(0)})|+|f(x^{(n)})|}$$

称为相对**平坦值**．算法在每个主步骤需要一个停止检验，一种这样的检验是看相对**平坦值**是否小．可以增加其他一些确实产生改进的检验，在算法的程序设计中，我们应该使 f 的求值次数最小，也应该避免利用指标的置换重新安排点使得 $f(x^{(0)})\geqslant f(x^{(1)})\geqslant\cdots\geqslant f(x^{(n)})$．事实上，只需要三个指标：最大的、次大的和最小的 $f(x^{(i)})$的指标.

除了 Nelder and Mead[1965]原文外，我们可以查询 Dennis and Woods[1987]以及 Torczon[1997]这些文章，不同的作者给出的算法的形式稍微有点差别．我们沿用了 Nelder 和 Mead 原来的描述．利用所谓的**增强下降**技巧的一种改进形式由 Tseng[1998]给出，也可见 Nazareth and Tseng[1998].

11.6 模拟退火法

已经提出这个方法并且有效地寻找困难函数的极小化，特别是当它们有许多真正的局部极小点的情况.

已知函数 $f:\mathbb{R}^n\to\mathbb{R}$，对$\mathbb{R}^n$ 中的任何 x，必须能计算 $f(x)$的值．我们打算求 f 的整体极小点，即求点 x^* 使对$\mathbb{R}^n$ 中的一切 x，有 $f(x^*)\leqslant f(x)$．换言之，$f(x^*)$等于 $\inf_{x\in\mathbb{R}^n}f(x)$．算法生成点 $x^{(1)}$，$x^{(2)}$，…的序列，并且希望 $\min_{j\leqslant k}f(x^{(j)})$收敛于 $\inf f(x)$（当 $k\to\infty$时).

假定 $x^{(k)}$已算出，描述导出 $x^{(k+1)}$的计算就足够了．在 $x^{(k)}$的一个大的邻域中生成一定数量

的随机点 $u^{(1)}$，$u^{(2)}$，…，$u^{(m)}$，对每一个点计算 f 的值．在序列中选取点 $u^{(1)}$，$u^{(2)}$，…，$u^{(m)}$ 之一为下一个点 $x^{(k+1)}$．选取指标 j 使得

$$f(u^{(j)}) = \min\{f(u^{(1)}), f(u^{(2)}), \cdots, f(u^{(m)})\}$$

723

如果 $f(u^{(j)})<f(x^{(k)})$，则取 $x^{(k+1)}=u^{(j)}$．否则，对每个 i，我们对 $u^{(i)}$ 用公式

$$p_i = e^{\alpha[f(x^{(k)})-f(u^{(i)})]} \qquad (1 \leqslant i \leqslant m)$$

指定一个概率 p_i，这里 α 是一个由代码的用户选择的正参数．我们用 m 个概率的和除每个 p_i 来规范化概率，即计算

$$S = \sum_{i=1}^{m} p_i$$

然后作一个替换

$$p_i \leftarrow p_i/S$$

最后，在点 $u^{(1)}$，$u^{(2)}$，…，$u^{(m)}$ 中作一个随机的选择，考虑到对它们已经指定了概率 p_i．随机地选取 $u^{(i)}$ 变为 $x^{(k+1)}$．

作这个随机选择最简单的方法是调用随机数生成元，在区间(0，1)中得到一个随机点 ξ．选取 i 作为使

$$\xi \leqslant p_1 + p_2 + \cdots + p_i$$

成立的第一个整数．因此，若 $\xi \leqslant p_1$，则设 $i=1$，$x^{(k+1)}=u^{(1)}$．若 $p_1<\xi\leqslant p_1+p_2$，则设 $i=2$，$x^{(k+1)}=u^{(2)}$，等等．

概率 p_i 的公式取自热力学理论，关于它的推导，读者可以查阅 Metropolis et al. [1953] 或 Otten and van Ginneken[1989]的原文．估计其他的函数可扮演这个角色．

复杂地选择 $x^{(k+1)}$ 的目的是离开真正的局部极小值．为此，算法必须偶尔地在当时的点中选择一个位于高处的点．因而存在一个机会使得后继的点可以开始移动到一个不同的局部极小点．

作一些小小的修正，算法可以用于一般的函数 $f: X\to\mathbb{R}$，其中 X 是任意集合．例如，在货郎担问题中，X 是一组整数{1，2，…，N}的一切置换集．所需要的全部是一个生成随机置换的过程以及对函数 f 求值的代码．

这个算法的 Fortran 程序可以在 Netlib 中找到：

http://www.netlib.org

11.7 遗传算法

由于需要防止从停止在一个真正的局部极小点上开始搜索，在最优化理论中一个新的进展导致一类称为**遗传算法**的方法．这些方法的思想来自于生物科学的观察，算法有一个随机的组成部分，迫使极小化搜索远离当时的局部极小值．

724

这类算法是由 Holland[1989]所描述的，并且关于这个主题的一本著作不久以后出版，见 Lawrence[1991]．这些算法只需要函数 f(典型地是定义在$\mathbb{R}^n$上)的值是可计算的——不需要导数．**货郎担问题**的遗传算法出现在 Karloy[1993]中．

通常为描述这样一个算法怎样工作，我们假设 f 是一个$\mathbb{R}^n$ 上的正函数，不知道 $f(x)$的极小值，但是它肯定大于或等于 0. 若缺乏任何较好的信息，则称 $f(x)$为 x 的**劣性**. 我们寻找一个最小劣性的 x，组合一对点产生一个或多个新点. 在这个过程中，低劣性的两个点比高劣性的点对具有较高的组成新点的概率. 在这个方法中点复制的概率反比于它的劣性. 一个新点由其他的两个点组合产生的方式与手头上的问题有关. 在算法的开始，利用随机数产生一组随机配置的点. 如果有充分的知识使这项工作值得去做的话，则随机点的分布可以预先规定. 因为组合的是原始对象总体的点对，所以在计算中存在大比例或大规模并行性的可能性：许多点对可以同时组合.

11.8 凸规划

这个术语表示当求极小化的函数是**凸的**并且求极小化的域是凸集时产生的特殊的最优化问题. 在 6.9 节中定义和讨论了凸集. 函数 f 的**凸性**是指每当 x 和 y 在 f 的定义域中且 $\alpha \geqslant 0$，$\beta \geqslant 0$和 $\alpha+\beta=1$ 时

$$f(\alpha x+\beta y) \leqslant \alpha f(x)+\beta f(y)$$

如果这样的一个函数是定义在$\mathbb{R}^n$ 的一个凸子集上，则它可以用下列形式的函数来逼近.

$$F(x)=\max_{1 \leqslant i \leqslant k}\left[\sum_{j=1}^{n} a_{ij} x_{j}-b_{i}\right]$$

类似地，凸域可以用下列形式的集合来近似表示

$$K=\{x: G(x) \leqslant 0\}$$

其中 G 是另一个像 F 那样同类型的凸函数，譬如说

$$G(x)=\max_{k<i \leqslant m}\left[\sum_{j=1}^{n} a_{ij} x_{j}-b_{i}\right]$$

725

相应地，我们考虑由函数对(F, G)描述的一个一般性问题，它指的是求服从于约束 $G(x)\leqslant 0$ 的 F 的极小点. 注意用上面的记号，倘若关键性的指标 k 已规定，则用矩阵 A 和向量 b 来完整地描述这个问题. 假定矩阵 A 满足 **Haar 条件**：A 的每个 n 个行向量组线性无关. 我们也假定凸集 K 非空. 定义 A 的行向量为：

$$A_i=(a_{i1}, a_{i2}, \cdots, a_{in}) \qquad (1 \leqslant i \leqslant m)$$

再定义**残差函数** r_i 为

$$r_i(x)=\sum_{j=1}^{n} a_{ij} x_j-b_i=\langle A_i, x\rangle-b_i \qquad (x \in \mathbb{R}^n, 1 \leqslant i \leqslant m)$$

在算法的每一步，给出一组指标 J，其中 $J \subset \{1, 2, \cdots, m\}$，它要求

1. J 恰好有 $n+1$ 个元素.
2. 至少 J 的一个元素在范围$\{1, 2, \cdots, k\}$中.
3. $\mathbb{R}^n$ 的原点位于行向量组$\{A_i: i \in J\}$的凸包中.

这个典型步骤如下. 首先，计算向量 $x \in \mathbb{R}^n$ 和实数 λ，如果 $i \in J$ 且 $i \leqslant k$，使得 $r_i(x)=\lambda$；如果 $i \in J$ 且 $i>k$，使得 $r_i(x)=0$. 这个计算需要解一组有 $n+1$ 个未知数(即 x 的分量和数 λ)的$n+1$ 个线性方程. 我们用手头上的 x 和 λ 计算 $F(x)$和 $G(x)$. 若 $G(x)\leqslant 0$ 和 $F(x)=\lambda$，则

算法停止，x 是解(即 $F(x)$ 达到极小值的 K 的点). 若 $G(x)>0$，则 $x\notin K$，选择一个指标 α 使得 $\alpha>k$ 且 $r_\alpha(x)>0$. 若 $G(x)\leqslant 0$ 和 $F(x)>\lambda$，则选择一个指标 α 使得 $\alpha\leqslant k$ 且 $r_\alpha(x)>\lambda$. 利用 6.9 节的**交换定理**，可求出一个指标 $\beta\in J$，该指标具有这样的性质，即 $\mathbb{R}^n$ 的原点位于 $\{A_i: i\in J'\}$ 的凸包中，其中 $J'=J\cup\{\alpha\}\setminus\{\beta\}$. 下一步从这个新的集合 J' 开始.

定理 1(有限终止定理)　*在有限步数之内，上面的算法产生一个点 $x\in K$，在这个点上 $F(x)$ 是一个极小值.*

这个定理及其证明可在 Goldstein[1966]中找到.

11.9　约束极小化

在前节中讨论了约束极小化的一种重要的情况，如何对待更一般的问题中不可利用域和函数的凸性的情况呢? 一个简单的方法提出了**罚函数**概念，这是一个把约束集外的点变得非常大的一个函数. 罚函数可以和目标函数相加，并且可以利用前面讨论的方法搜索和函数的无约束极小值.　726

更特殊些，假定求极小值的目标函数是 f，约束条件为 $g(x)\leqslant 0$. 换言之，在集合 $K=\{x: g(x)\leqslant 0\}$ 上求 $f(x)$ 的极小值. 约束问题等价于 $f+g^*$ 的无约束极小化，其中罚函数依赖于 $g(x)>0$ 或 $g(x)\leqslant 0$ 满足条件 $g^*(x)=\infty$ 或 $g^*(x)=0$. 如人们可能猜测的那样，这个人为的函数 g^* 太粗糙因而在数值过程中并不实用. 代之，当 $g(x)$ 上升到临界值 0 以上时，我们应该使用一个迅速上升的光滑函数. 例如，可使用替代的 $g^*(x)=\exp(g(x))-1$.

另外可使用的罚函数是 $g^*(x)=\max\{0, \alpha g(x)\}$，其中 α 是一个由专业人员选择的正参数.

11.10　帕雷托最优化

术语**帕雷托最优化**应用于一组有限个实值函数 $f_1, f_2, \cdots, f_n$ 的同时极小化，可以设想好几个最优化准则. 例如，我们可求

$$\sum_{i=1}^{n} f_i(x)$$

的极小值或者求

$$\max_i f_i(x)$$

的极小值. 这两种方法将直接导致一个实值函数的极小化问题，这类问题在本章的前面几节已作了讨论.

在帕雷托最优化中，我们寻找一个点 x^* 使得没有其他的点 y 满足不等式 $f_i(y)<f_i(x^*)$，$1\leqslant i\leqslant n$. 这里这样的 x^* 称为已知的函数族的**帕雷托最优点**.

再次需要集合凸性以及 11.8 节中定义的函数的凸性的概念. 下列定理给出帕雷托最优点的必要条件，本定理及下一个定理取自 Aubin[1998].

定理 1(帕雷托极小必要条件定理)　*设 $f_1, f_2, \cdots, f_n$ 是定义在某个向量空间中的一个凸集上的凸实值函数，若 x^* 是这个函数族的一个帕雷托极小点，则存在一个非零非负向量 $(\lambda_1, \lambda_2, \cdots, \lambda_n)$ 使得 x^* 极小化 $\sum_{i=1}^{n}\lambda_i f_i(x)$.*　727

证明 设函数的凸域是 X，用

$$f(x)=[f_1(x),f_2(x),\cdots,f_n(x)]$$

定义 $f:X\to\mathbb{R}^n$. 在$\mathbb{R}^n$中，定义 $u\leqslant v$ 意味着对一切 i，$u_i\leqslant v_i$，而 $u<v$ 意味着对一切 i，$u_i<v_i$. 我们也取

$$K=\{u\in\mathbb{R}^n:u>f(x),\text{对某个 } x\in X\}$$

我们证明集合 K 是凸的. 假设 u，$v\in K$，则对某个适当的 x，$y\in X$，有 $u>f(x)$和 $v>f(y)$. 若 $0\leqslant\lambda\leqslant1$ 及 $\theta=1-\lambda$，则由每个函数 f_i 的凸性，有

$$\lambda u_i+\theta v_i>\lambda f_i(x)+\theta f_i(y)\geqslant f_i(\lambda x+\theta y)=f_i(\omega)$$

其中，$w=\lambda x+\theta y$. 由 X 的凸性，w 是 X 的一个点. 刚才所写的式子表明 $\lambda u+\theta v>f(w)$，因此，$\lambda u+\theta v\in K$ 并且 K 是凸的.

现在证明定理. 假设 x^* 是函数族 f_i 的一个帕雷托极小点，我们断言 $f(x^*)$不在 K 中. 事实上，如果 $f(x^*)$属于 K，则对 X 中的某个 x，一定有 $f(x^*)>f(x)$，直接与帕雷托最优点的含意矛盾. 应用 10.1 节中的分离定理 I，在$\mathbb{R}^n$ 中得到一个从 K 中分离 $f(x^*)$的超平面. 于是，对某个非零向量 $h\in\mathbb{R}^n$ 和 K 中的一切 u，$\langle h,f(x^*)\rangle\leqslant\langle h,u\rangle$. 因为 K 包含一切形如 $f(x)+p$ 的向量，其中 $x\in X$，$p\in\mathbb{R}^n$ 且 $p>0$，所以有

$$\langle h,f(x^*)\rangle\leqslant\langle h,f(x)\rangle+\langle h,p\rangle$$

考虑非常大量的 p，我们得到 $h\geqslant0$，取变量 p 的下确界给出

$$\langle h,f(x^*)\rangle\leqslant\langle h,f(x)\rangle$$

这断言 x^* 在 X 上极小化 $\sum_{i=1}^{n}h_if_i$. ■

定理 2(帕雷托极小充分条件定理) 若点 x^* 极小化 $\max f_i(x)$ 或 $\sum_{i=1}^{n}\theta_if_i(x)$，其中 $\theta_i\geqslant0$ 且 $\sum_{i=1}^{n}\theta_i>0$，则 x^* 是函数系 $f_1,f_2,\cdots,f_n$ 的一个帕雷托最优点.

证明 假设 x^* 不是一个帕雷托最优点，则存在一个点 y 使得对一切 i，$f_i(y)<f_i(x^*)$. 因此，$\max_i f_i(y)<\max_i f_i(x^*)$. 从而，$x^*$ 不极小化 $\max f_i(x)$，对$\sum\theta_if_i(x)$情况可类似地证明. ■

例 1 设 $f_1(x)=x^2$，$f_2(x)=(1-x)^2$，其中 $x\in\mathbb{R}$，点 $x^*=\frac{1}{2}$是函数系$\{f_1,f_2\}$的一个帕雷托极小点. 确实，由 $f_i(y)<f_i(x^*)$可得到结论 $|y|<\frac{1}{2}$和 $|y-1|<\frac{1}{2}$. 这些 y 的不等式是互斥的. 定理 1 证明中提到的集合 K 是

728

$$K=\{u\in\mathbb{R}^2:u_1>x^2\text{ 和 }u_2>(x-1)^2\text{对某个 }x\in X\}$$

方程是 $u_1+u_2=\frac{1}{2}$的超平面把 K 和 $f(x^*)=\left[\frac{1}{2},\frac{1}{2}\right]$分离. 因此，如定理 2 的证明中那样，点 x^* 实际上极小化函数 f_1+f_2. 因为 x^* 也极小化 $\max_i f_i$，所以此例同样说明定理 2. ■

习题 11.10

729

证明：若函数 f_i 是凸的，则当 $\lambda_i\geqslant0$ 时，$\max f_i$ 和 $\sum_{i=1}^{n}\lambda_if_i$ 也是凸的.

附录A　数学软件一览

全世界有大量的数学软件可用，并且其数量日益增加. 为了获得最新的信息，可浏览因特网上的万维网. 在因特网上，可以搜索到用户所关心的特定应用领域中的常用数学软件.

把数学软件分成如下3类是有益的：1)不受版权限制，2)可免费获得(某些功能限制使用)，3)有专利权(需要特许契约). 不受版权限制有一个特殊的法律意义，指的是任何用途都允许，包括修改、转售等等. 通过因特网，人们可以下载不受版权限制或可免费获得的软件，并且在许多情况下，能获得商业软件包的免费演示版. 因特网上的某些软件是有使用限制的，例如作者的版权，只以研究与教学为目的的使用不受作者的版权限制(类似的例子是ACM Collected Algorithms的代码，它受制于ACM软件版权和特许契约). 另一方面，专有软件必须从销售软件的开发公司或计算机商店购买或租借. 是否在软件能使用之前付费不只是着眼于软件的分类. 例如，仔细考虑下面这些看上去矛盾的例子：(1)付费买不受版权限制的软件(比如，CD-ROM上的Netlib)，(2)付费买容易得到的软件——你能免费下载某些软件，但是如果用于非教育目的，那么你最好递送一张支票，(3)使用免费赠送的软件，但它是有专利的，在你得到它之前需要签署一个特许契约.

在这个附录中，我们对某些数学软件作一些简要介绍并且指出其在因特网上的万维网(www)地址，用户可以在这些地址上发现补充信息和一些软件. 我们还要提到若干个系统，这些系统有助于搜索求解特定问题的数学软件. 因为软件的开发速度如此之快，以至于任何一份软件清单都会很快过时，所以这里提供的不是一张全面的清单.

因为数学软件是用各种程序设计语言编写的并且适用于多种计算机体系结构，所以怎样更有条理地介绍这些软件是一件困难的事情. 我们受到Lozier and Olver[1994]有关数学特殊函数的精彩评述的影响. 他们按照软件包、中间程序库、综合程序库和交互系统等类别来组织软件. 首先，软件包包含有一个或多个子程序，这些子程序是用来求解数值数学的一个子域中的特殊问题. 其次，中间程序库通常是适用于小型计算机或PC机的子程序. 有些库包含用一种或多种计算机语言编写的实用数学函数的集成. 它们有可能是由计算机设备制造商或编译器的开发者编写. 另外，有些适用于PC机的中间程序库是通用数学程序库的子库. 第三，综合程序库包含的子程序已被吸收到许多特征统一的高质量软件中，比如，统一的文本、统一的使用方式和差错处理条件. 最后，交互系统至少是带有一组功能强大的键盘命令的交互式计算机环境，以便用户能避免编译-连接-运行的循环. 数学软件通常要经历一个改良过程，从期刊或报告中介绍的原始思想到被算法科学界普遍接受. 最终，把算法合并到大型数学程序库或成熟的交互式计算机系统中去.

我们先提及在因特网上寻找常用数学软件的某些搜索系统，然后给出与数学软件研究和开发相关的一般信息. 最后，讨论若干综合程序库和交互系统，以及一些中间程序库或数学软件包的例子.

A.1 搜索系统

A.1.1 Guide to Available Mathematical Software (GAMS)

这是一个大型在线式的附录前后参照索引的虚拟数学及统计软件资料档案库，它适用于计算科学和工程学. 它是国家标准技术研究所(NIST)的一个开发项目. 这个项目给科学家和工程师提供可重复使用的计算机软件的更便捷的入口. 用户以问题决策树为向导来搜索适当的软件以便于解决特殊的问题，或者搜索包/模块名. GAMS 不提供实体的资料档案库，而只提供显而易见的由 NIST 等来维护的多重资料档案库的入口. 用户可以在

http://gams.nist.gov[⊖]

上获得摘要、文档和源代码. GAMS 具有附加搜索特征的一个 Java 界面可以在

http://gams.nist.gov/HotGAMS

上找到. 有关数学软件的其他一些信息源可以在

http://gams.nist.gov/OtherSources.html

732
上找到，其中包括目录、期刊、免费的以及来自教育和其他软件供应商的可用程序包.

A.1.2 Netlib

这是数学软件、文档(论文、报告等)、数据库(地址清单(电子邮件和常规信件)、会议、性能数据等)和其他实用的数学信息的资料档案库. 用户可以执行一个关键字搜索来获得数学软件以及搜索以前每周的 NA-Digest 在线通讯稿的议题. 这是一个为数值分析工作者和其他研究人员的社团提供服务而开发的系统，用户可以通过

http://www.netlib.org

进入这个系统. 另外，用户也可以从

http://netlib.org/na-net/

进入 NA-Net 系统. 它服务于科学计算界，为其成员提供 e-mail 数据库、有关用户信息的名录服务和 NA-Digest——每周热点主题的论文汇集. 其他一些相关的成果如下：

- HPC-Netlib 是 Netlib 的高性能分支，可以从

  http://www.nhse.org/hpc-netlib/

  上得到. 它提供有关高性能数学软件(不管是用于研究还是商业)的资料，并且还给出诸如软件选择和执行的路标.
- Matrix Market 可以从

  http://math.nist.gov/MatrixMarket

  进入，它提供一个进入数值线性代数算法比较研究中使用的测试数据资料档案库的便利入口.
- PTLib 是有关并行系统的高质量软件和工具的资源. 它在

  http://www.nhse.org/ptlib

⊖ 因特网上的信息是变化的，这里列出的网址未必都能用. ——编辑注

上可以找到，它是国家高性能计算和通信软件交换站的一部分.

A.2 一般信息

A.2.1 Conferences

这是有关会议信息的公告，可以在因特网上获得相关的信息. 比如，从

http://www.netlib.org/confdb/Conferences.html

可以进入 Netlib Conferences Database，它包含即将召开的会议、讲演、以及与数学和计算机科学领域相关的其他会议的信息. 此外，Atlas Mathematical Conferences Abstracts 在

http://at.yorku.ca/amca

733

上为全世界数学家提供会议公告和会议摘要. 多数会议公告可从下面列举的科学社团和协会直接得到.

A.2.2 Graduate

这是涉及数值分析和科学计算的程序，遍布世界各地，比如得克萨斯大学奥斯汀分校的计算与应用数学(CAM)程序的细节可以在

http://www.ticam.utexas.edu/cam

上得到. 在万维网上可以获得许多有关其他研究生程序的资料.

A.2.3 Homepages

这是由各种研究兴趣团体建立的. 例如：

- 区间运算的最近的研究进展：

http://cs.utep.edu/interval-comp/main.html

- 最优化软件的决策树：

http://plato.la.asu.edu/guide.html

个人的主页，比如本书作者的主页是：

- David Kincaid 主页

http://www.cs.utexas.edu/users/kincaid

- Ward Cheney 主页

http://www.ma.utexas.edu/users/cheney

多数专业社团和协会都有主页，比如下面列出的社团和协会.

A.2.4 Journals

这是为了传播最新研究进展和相应的算法与数学软件而出版的. 本书的参考书目包含了许多主要的数值分析期刊的清单. 其中有些期刊读者可以在因特网上获得其目录甚至论文的全文. 出版研究期刊和传播数学软件的组织有：

- Society for Industrial and Applied Mathematics (SIAM)(工业与应用数学协会)：

http://www.siam.org

- Association for Computing Machinery (ACM)(计算机协会)：

734

http://www.acm.org

- American Mathematical Society (AMS)(美国数学学会)：

http://www.ams.org

- Mathematical Association of America (MAA)(美国数学协会)：

http://www.maa.org

- International Association for Mathematics and Computers in Simulations (IMACS)(国际数学和计算机仿真协会)：

http://www.imacs_online.org

- Journal of Approximation Theory (JAT)：

http://www.math.ohio-state.edu/JAT

例如，《ACM Transactions on Mathematical Software(TOMS)》出版审定过的论文和计算机程序/包. ACM 算法的指导原则要求软件是具有适当文档独立成套的，有一个样板输出的测试程序，以及能较好地将其移植到各种计算机上. TOMS 的主页

http://www.acm.org/toms

上有算法论文目录的搜索表和有关软件链接的搜索表. 从 ACM、Netlib 和许多其他源点可以得到软件. 用户能从

http://www.acm.org/pubs/calgo

上的 ACM Collected Algorithms 获得资料. 分类系统是用来编制算法索引的，而且这些算法的数据库得到很好地维护. ACM Digital Library 是文献信息，引文和论文全文的源点；它可以在

http://acm.org/dl

上得到. 文献数据库对注册访问者是免费的，进入全文数据库是按订阅服务次数计费的. 许多新的和经典的数学著作可以在

http://siam.org/books/

上得到. 数学期刊，比如《SIAM Journal of Numerical Analysis》，《SIAM Journal of Scientific Computing》和其他许多期刊，可以在

http://siam.org/journals/journals.htm

上得到. 其他一些期刊为特殊的科学学科中的软件及其相关信息的交换提供平台. 例如，期刊《Computer Physics Communications》发表有关物理学和物理化学的计算方面的论文并伴有审定过的计算机程序. 期刊《Applied Statistics》发表统计计算的文献资料并伴有审定过的统计软件. 最近，已出现有关数值分析算法进展的专门论文的纯电子期刊. 因特网上得到的这些期刊的论文可以本地打印. 这类期刊之一是在

735

http://etna.mcs.kent.edu

上的《Electronic Transactions on Numerical Analysis》.

A.2.5 Mathematics Archive WWW Server

这是提供进入各种各样数学资源的有组织的因特网入口. 重点是那些用于传授数学和教育软件的材料. 它可以在

http://archives.math.utk.edu

上得到.

A.2.6 Mathematics Information Servers

这是万维网上一份详尽的数学服务器清单，涉及许多院系、电子期刊、预印论文的原始资料以及有关数学软件的资料. 它可以在

http://www.math.psu.edu/MathLists

上得到.

A.2.7 News Groups

用于以无节制论坛的方式张贴邮件，这些邮件讨论的内容广泛而形式为问答式. 关于数学软件的一些 USENET 新闻组是

sci.math.num-analysis
sci.math.research
sci.math.symbolic

A.2.8 Newsletters

在因特网上发布公告和有关数学软件的一般信息. 范例有：

- NA-Digest，数值分析在线通信：

  http://www.netlib.org/na-digest-html/index.html

- MGNet-Digest，多重网格在线通信：

  http://www.mgnet.org/mgnet-digest.html

- AT-Net，逼近论网络与在线公报：

  http://www.uni-giessen.de/www-Numerische-Mathematik/at-net

A.2.9 Research

世界各地都有数值分析与科学计算的中心和研究所. 比如，在得克萨斯大学奥斯汀分校网站

http://www.utexas.edu

中的得克萨斯计算与应用数学研究所(TICAM)：

http://www.ticam.utexas.edu

和数值分析中心（CNA)：

http://www.ma.utexas.edu/CNA

在美国，大量的数学软件可从国家实验室(Argonne[ANL]、Fermilab、Lawrence Berkeley[LBNL]、Lawrence Livermore[LLNL]、Los Alamos[LANL]、Oak Ridge[ORNL]、Pacific Northwest、Sandia 和 Stanford Linear Accelerator 等)、国家超级计算机中心(圣迭戈超级计算机中心[SDSC]、在劳伦斯·伯克利的国家能量超级计算机中心 [NESC]等)和各种政府机构(国家标准技术研究所[NIST]等)获得. 用户可以直接链接到所有的国家实验室、超级计算机中心和高性能研究中心.

- 高性能计算中心清单：

http://www.nhse.org/univ_hpcc.html

- 美国政府实验室和机构清单：

http://www.nhse.org/gov_sites.html

A.2.10 Textbooks

教科书与相关的数学软件能广泛地得到. 许多数值分析和数值方法的教科书与相关的软件一起供给，或者已把软件与书合在了一起. 在这些书中，对每个问题所涉及的领域人们能找到分析数学的一般讨论、算法的表述和以一种或几种计算机语言编写的计算机程序算法的实际实现. 算法可能列在课本中，也可以购买软盘或者免费获得软件，感兴趣的读者可从因特网上下载软件. 例如支撑本书的软件可以在下列网站：

http://www.ma.utexas.edu/CNA/NA3

或

ftp://ftp.brookscole.com/dir/brookscole/Mathematics/Texts_by_Authors/Kincaid_Cheney

中得到. 在那里，用户可以下载样板计算机程序，这些程序或者用各种计算机程序设计语言编写，或者使用高级数学软件系统. 同时，还能得到本书最新的勘误表. 此外，支撑初级教科书《Numerical Mathematics and Computing，4th Edition》(Cheney 和 Kincaid 编著)的软件可在下列网站得到：

736～737

http://www.ma.utexas.edu/CNA/NMC4

A.3 综合程序库

综合程序库是以统一的风格和高质量及稳健的规格写成的大型数学程序的集成. 这里列举一些综合程序库.

A.3.1 CERN 程序库

这是由欧洲粒子物理实验室提供的. 这个程序库主要是支撑高能物理研究，但是它包含许多一般数学上使用的程序. 这个程序库经销给外单位时带有某些限制. 它在

http://consult.cern.ch

上可以得到.

A.3.2 CMLIB

这是国家科技研究所 (NIST)的核心数学程序库. 它是一个大约有 750 个高质量的不受版权限制的 Fortran 子程序的集成，而且这些子程序很容易移植. 这个库中的子程序解决数学和统计学中许多标准的问题，它包含通常在其他地方可得到的软件程序，例如，BLAS、EISPACK、FISHPACK、FCNPACK、FITPACK、LINPACK 和 QUADPACK. 可访问

http://gams.nist.gov

A.3.3 ESSL/PESSL

这是一个 IBM 计算机上使用的工程和科学子程序库. 它是一个适用于在各类 IBM 计算机上求解科学和工程应用问题的达到目前最新水平的 450 个以上数学程序的集成. 程序库对特殊

的 IBM 计算机体系结构(例如工作站和并行计算机)作了协调. 它可以调用以 Fortran, C 或 C++ 编写的应用程序. PESSL 是 IBM 工程和科学子程序库(ESSL)的并行实现. 它由在 IBM 计算机系统及 IBM 工作站群中支持并行处理的可升级的数学程序组成. 并行 ESSL 支持单程序多重数据(SPMD)程序设计模型, 它或者利用信息传递界面(MPI)信号处理程序库或者利用 MPI 插入程序库. 请访问

http://www.ibm.com

A.3.4 IMSL 程序库

这是由 Visual Numerics 有限公司开发的 C 代码或 Fortran 代码数值和图像程序库. 这个程序库包含一个大型的子程序和函数子程序库(500 个以上)的集成. 它提供进入数学和统计学中数值方法高质量实现的入口. 这些程序库经过 30 多年演变而来, 它们在各类计算机平台上使用都是有效的, 它们的子集在 PC 机上使用也是有效的. Visual Numerics 中其他可以利用的数学软件产品是 PV-WAVE 和 Stanford Graphics. 请访问

http://www.vni.com

738

A.3.5 LibSci

这是为适用于 Cray 计算机系统开发的通用的数学和科学程序库. 例如, 它包含线性代数、快速傅里叶变换、滤波、打包/解包和向量聚集/分散. 这个程序库在

http://www.netlib.org/scilib

上可以得到.

A.3.6 NAG 程序库

这是科学家、工程师、研究员和软件开发人员的 Fortran77/90 代码或 C 代码数值和统计程序库, 涉及数学、统计和最优化的应用程序. 它由 Numerical Algorithms Group(NAG)开发的, 这些软件库具有综合的数值性能, 是数值专家和统计专家协作编写的代码. 程序库的最大版本具有 1000 个以上的程序, 其中若干软件经过 20 多年演变而来. 具有目前最新水平的产品在从 PC 机到超级计算机的 80 多个计算机平台上都具有稳健的性能. 此外, NAG 建立了第一个完整的标准 Fortran 90 编译器, 并且根据 NAG, 人们可利用计算机代数系统 Axiom. 信息在

http://www.nag.com

上可以得到.

A.3.7 SLATEC

这是由能源部(DOE)和能源科技中心经售的一个大型 Fortran 数学子程序集成. 用户可以从

http://www.energy.gov

上或者从 Netlib 的

http://www.netlib.org/slatec

上得到. 它具有可移植性、良好的数值技术、良好的文档、稳健性和质量保证等特点. 这个程

序库最初是为政府研究实验室联盟的超级计算机提供可移植的、私人拥有的数学软件. 原缩写词代表涉及的国家实验室(Sandia、Los Alamos 和 Air Force Technical Exchange Committee). 后来，程序库委员会接纳了另外三个国家实验室(Lawrence Livermore、Oak Ridge 和 Sandia Livermore)以及在 Lawrence Livermore 的国家能源超级计算机中心和国家标准和技术协会.

除了上面列举的程序库以外，其他一些一般的数学软件程序库是 NSWC、NUMAL、NUMPAC、PORT、Scientific Desk 和 VECLIB 等.

A.4 交互系统

一般来说，一个完整的交互数学软件系统含有一组强大的用户进入计算机终端或工作站的指令. 在屏幕上显示直接的响应. 它可能是一个计算(数值的或符号的)或是一个视觉显示，例如，一个图表或图形. 因为不存在编译-链接-执行循环，所以减轻了程序设计的负担. 通过定制指令集或图标可以扩展交互系统的性能. 交互系统可把非数值工作和数值计算结合起来. 看来在一个交互式计算机环境内做图像计算和符号计算似乎是最好的. 由于这个原因，一个新趋势是把数值计算与符号计算和图像可视化结合成为一个完全的交互系统.

739

计算机代数系统具有在数值数学中有用的特殊性能. 一种这样的特性是任意精度或多精度浮点运算. 一般来说，程序设计语言使用计算机硬件做计算机运算使得精度是固定的. 计算机代数系统的主要目的是作精确的数学运算，具有浮点运算是它的另一个性能. 尽管如此，我们得到的一个意外收获是能够在这个系统内执行任意精度的浮点运算. 除非用户另有说明，符号系统通常避免引入不精确结果的求值以及计算符号表达式(数表示为具有任意长的分子和分母的有理分数或表示为符号). 用户可以要求以任意长的精度执行数的浮点运算.

下面是一些交互数学软件系统的样例.

A.4.1 CPLEX

这是一个用于求解线性规划问题的软件包，它包括整数、混合整数以及网络线性规划问题. CPLEX 可作为一个交互程序或作为一个可调用的子程序库使用. CPLEX 交互问题求解程序允许用户进入、修改和求解来自计算机终端的问题. 当一次求解一个问题或者设计一个解法时这是特别有用的. 可调用的程序库提供访问 CPLEX 最优化、公用程序、问题修正、查询和直接来自 C 或 Fortran 程序的 I/O 程序的入口. 请访问

http://www.cplex.com

A.4.2 HiQ

这是一个基于对象的数值分析和数据可视化软件包. HiQ 利用兼有工作表界面、交互分析、数据可视化、一个扩充的数学程序库和一个脚本程序设计语言的一套方法解决数学、科学和工程问题. HiQ 对 Macintosh 和 Power Macintosh 计算机是一个交互的问题解决环境. 它可以在

http://www.ni.com

上得到.

A.4.3 Maple

这是一个包含符号的、数字的、图像的和程序设计的性能的交互符号计算系统. Maple 执行方程求解、线性代数、微积分和复分析，并且具有实质上无限的精度. Maple 是作为与符号计算有关的计算机代数操作的一个交互系统开发的，但是加入了许许多多扩充的性能. 从个人计算机和工作站到超级计算机等各种类型的计算机它都可使用. 它可以在

http://www.maplesoft.com

上得到.

740

A.4.4 Mathematica

这是一个用于数值计算、符号计算和图像计算以及可视化的交互软件系统. 用户可以把 Mathematica 的输出信息(计算、图像和动画等)和原始文本在 Mathematica 系统中完整地结合起来，作为用于展示或技术报告的完整的备用电子文件. 可得到多种 Mathematica 应用程序库. MathLink 是一个通信协议，它允许 Mathematica 和其他计算机程序包，如 Matlab 或 Excel 之间交换信息. 基于模式匹配的程序设计语言可用于扩充 Mathematica 系统的性能. 从 PC 机到科学工作站，再到大规模的特大型计算机，在 20 多个平台上都可使用 Mathematica. 请访问

http://www.mathematica.com

A.4.5 Matlab

这是一个提供计算、可视化和特殊应用软件工具箱的计算环境. 矩阵记号用于产生带有一组涉及数值计算的标准算法的内置指令的矩阵实验室. 可用一个面向矩阵的语言进行大规模的计算和数据分析. 交互 2D 和 3D 图像的性能对数据的分析、变换和可视化是有用的. Matlab 可动态地链接 C 或 Fortran 程序. 对 PC 机、工作站和超级计算机等各种类型的计算机都可使用该程序包. Matlab 具有大于 20 个的工具箱，可用于特殊的应用，例如，信号和图像处理、控制系统设计、频域识别、稳健控制设计、数学、统计、数据分析、神经网络和模糊逻辑、最优化以及样条. 此外，用 Maple V 带工具箱界面也可作符号计算. 它可以在

http://www.mathworks.com

上得到.

A.4.6 Octave

这是一种高级语言，通常和 Matlab 兼容，并且主要用于数值计算. 它为数值求解线性和非线性问题以及执行其他数值实验提供指令行界面. Octave 中包含大量的工具，这些工具用于求解通常的数值线性代数问题和求非线性方程求根、求常用函数的积分、处理多项式以及求常微分方程和微分代数方程的积分. 借助于以 Octave 自己的语言编写的用户定义的函数或利用以不同程序设计语言编写的动态加载的模块，可以扩充和定制 Octave. GNU Octave 是一个免费的软件，根据免费软件基金会出版的 GNU 公众许可证，它可以重新分配或修改. 它在

http://www.octave.org

上可以得到.

A.4.7 REDUCE

这是为数学家、科学家和工程师感兴趣的普通代数计算设计的一个交互计算机代数系统. [741] 尽管对那些能用手工处理的问题 REDUCE 可作为一台代数计算器使用，但是它的主要目的是用来处理那些只能用计算机计算的问题. 请访问

http://www.uni-koeln.de/REDUCE/

还有一些不同于列举在此的某些符号代数计算机系统，例如 Axiom、Derive、GANITH、Macsyma、Magma、Mathcad、Milo、MuPAD、Pari、Schur 和 SymbMath.

A.5 中间程序库和软件程序包

可利用许多中间程序库和软件程序包求解各类数学问题. 一些主要的不受版权限制的软件样例如下：

A.5.1 FITPACK

这是一个在张力作用下利用样条执行曲线拟合和曲面拟合的数学程序库. 其特色包括具有各种维数的基和目标空间、带有明显极点的近似数据和近似积分、导数以及近似曲线的长度. 它在

http://www.netlib.org/fitpack

上可以得到.

A.5.2 ITPACKV/NSPCG

这是利用迭代法求解大型稀疏线性方程组的子程序包. 软件的特色包括自动参数估计和实现精确的终止准则. 这些程序包可用于求解椭圆偏微分方程的有限差分或有限元模型形成的线性方程组. ITPACKV 是具有基本迭代格式的向量化软件包. 这些格式为雅可比、逐次超松弛(SOR)、对称 SOR 以及用切比雪夫或共轭梯度法加速收敛的约化方程组. 在 NSPCG 的程序中也涉及各种加速技术，包括共轭梯度、切比雪夫加速和非对称方程组带预处理(或基本迭代法)的广义共轭梯度法. 这些程序包是作为得克萨斯大学奥斯汀分校数值分析中心的 ITPACK 项目的一部分开发的. 请访问

http://www.ma.utexas.edu/CNA

A.5.3 LAPACK

这是包含求解联立线性方程组、线性方程组的最小二乘解、特征值问题和奇异值问题的程序. 也提供有关的矩阵分解(LU、楚列斯基、QR、SVD、舒尔和广义舒尔)，它们是与诸如舒尔分解重新排序和估计条件数的计算有关的. 处理稠密的和带状的矩阵，但不处理一般 [742] 的稀疏矩阵. 在所有的领域中，对实阵和复阵在单精度和双精度两方面提供相似的功能. 请访问

http://netlib2.cs.utk.edu/lapack/index.html

A.5.4 MINPACK

这是一个求解非线性方程组和非线性最小二乘问题的程序库. 我们可选择使用自动地计算雅可比矩阵的程序也可为雅可比矩阵提供程序. 请访问

http://www-fp.mcs.anl.gov/otc/Guide/SoftwareGuide/index.html

A.5.5 ODEPACK

这是求解已知初值的常微分方程的子程序集成．这些程序对用线方法求解不定常偏微分方程时形成的常微分方程的求解特别有用．它们是由 Lawrence Livermore 国家实验室(LLNL)所开发的，这个程序包的主页是

http://www.llnl.gov/CASC/odepack

A.5.6 PDE2D

这是在一般的二维区域和三维盒体内求解线性和非线性不定常、定态及特征值偏微分方程组的一个有限元代码．它能处理诸如弹性、扩散、热传导、势能和流体力学等领域中的问题．它具有一个交互接口和大量的图像输出能力．它在

http://members.aol.com/pde2d/

上可以得到．

A.5.7 PETSc

这是求解偏微分方程及相关问题的科学计算的一个便携式可扩充的工具箱．它是一套创建大范围应用的数据结构和程序，这些应用由并行的线性和非线性方程求解程序和无约束最小化模块所组成．Argonne 国家实验室(ANL)支持这个项目并且是主要的发布站点：

http://www-fp.mcs.anl.gov/petsc

A.5.8 ScaLAPACK

这是由几个公共机构(Oak Ridge 国家实验室，莱斯大学，加州大学伯克利分校和洛杉矶分校，伊利诺伊大学，田纳西大学 Knoxville 分校)协作完成的成果．它由四个部分组成：稠密带状矩阵软件(ScaLAPACK)、大型稀疏特征值软件(PARPACK 和 ARPACK)、直接法求解稀疏方程组软件(CAPSS 和 MFACT)和迭代求解大型稀疏方程组的预处理求解程序(ParPre)．它在

http://netlib2.cs.utk.edu/scalapack/index.html

上可以得到．

743

参考文献

缩写

ACM	*Association for Computing Machinery*
ACM-COM	*ACM Communications*
ACM-J	*ACM Journal*
ACM-TOMS	*ACM Transactions on Mathematical Software*
AMM	*American Mathematical Monthly*
AMS	*American Mathematical Society*
AMS-B	*AMS Bulletin*
AMS-T	*AMS Transactions*
AN	*Acta Numerica*
ANM	*Applied Numerical Mathematics (IMACS)*
AMC	*Applied Mathematics and Computation*
ANSI	*American National Standards Institute, Inc.*
BIT	*BIT Numerical Mathematics*
CA	*Constructive Approximation*
CJ	*Computing Journal*
CANM	*Communications on Applied Numerical Methods*
CMP	*Communications in Mathematical Physics*
CPAM	*Communications on Pure and Applied Mathematics*
ETNA	*Electronic Transactions on Numerical Analysis*
IEEE	*Institute of Electrical and Electronic Engineers*
IEEE-TAE	*IEEE Transactions on Audio and Electroacoustics*
IEEE-TC	*IEEE Transactions on Computing*
IJNME	*International Journal for Numerical Methods in Engineering*
IMACS	*International Association for Mathematics and Computers in Simulation*
IMA-JNA	*Institute for Mathematics and Its Applications, Journal of Numerical Analysis*
IU-JM	*Indiana University Journal of Mathematics*
JAT	*Journal of Approximation Theory*
JCAM	*Journal of Computational and Applied Mathematics*
JCP	*Journal of Computational Physics*
JMP	*Journal of Mathematical Physics*
JR-NBS	*Journal of Research in the National Bureau of Standards*
LAA	*Linear Algebra and its Applications*
LNCS	*Lecture Notes in Computer Science*
LNM	*Lecture Notes in Mathematics*
MAA	*Mathematical Association of America*
MI	*Mathematics Intelligencer*
MOC	*Mathematics of Computation*
MP	*Mathematical Programming*
NM	*Numerische Mathematik*
SA	*Scientific American*
SIAM	*Society for Industrial and Applied Mathematics*

SIAM-MAA	*SIAM Journal on Matrix Analysis and Applications*
SIAM-NA	*SIAM Journal on Numerical Analysis*
SIAM-REV	*SIAM Review*
SIAM-SSC	*SIAM Journal on Scientific and Statistical Computing*
ZAMM	*Zeitschrift für Angewandte Mathematik und Mechanik*
ZAMP	*Zeitschrift für Angewandte Mathematik und Physik*

Abramowitz, M., and I. A. Stegun. 1956. Abscissas and weights for Gaussian quadratures of high order. *JR-NBS* **56**, 35–37.

Abramowitz, M., and I. A. Stegun (eds.) 1964. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. National Bureau of Standards. (Reprinted New York: Dover, 1965.)

Acton, F. S. 1959. *Analysis of Straight-Line Data*. New York: Wiley. (Reprinted New York: Dover, 1966.)

Ahlfors, L. V. 1966. *Complex Analysis*. New York: McGraw-Hill.

Aho, A., J. Hopcroft, and J. Ullman. 1974. *The Design and Analysis of Computer Algorithms*. Reading, MA: Addison-Wesley.

Aiken, R. C. (ed.). 1985. *Stiff Computation*. New York: Oxford University Press.

Alefeld, G., and R. Grigorieff (eds.). 1980. *Fundamentals of Numerical Computation*. Berlin: Springer.

Alefeld, G., and J. Herzberger. 1983. *Introduction to Interval Computations*. New York: Academic Press.

Alexander, J. C., and J. A. Yorke. 1978. The homotopy continuation method: Numerically implemented topological procedures. *AMS-T* **242**, 271–284.

Allgower, E., and K. Georg. 1980. Simplicial and continuation methods for approximating fixed points and solutions to systems of equations. *SIAM-REV* **22**, 28–85.

Allgower, E., and K. Georg. 1990. *Numerical Continuation Methods*. New York: Springer-Verlag.

Allgower, E. L., K. Glasshoff, and H.-O. Peitgen (eds.). 1981. *Numerical Solution of Nonlinear Equations*: *LNM* **878**. New York: Springer-Verlag.

Ames, W. F. 1977. *Numerical Methods for Partial Differential Equations*. New York: Academic Press.

Anderson, E., Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen. 1995. *LAPACK Users' Guide - Release 2.0*. Philadelphia: SIAM. To view online version, use the URL address: `http://www.netlib.org/lapack/lug/lapack_lug.html`

ANSI/IEEE. 1985. IEEE standard for binary floating-point arithmetic. ANSI/IEEE Std. 754–1985. New York: IEEE.

ANSI/IEEE. 1987. A radix-independent standard for floating-point arithmetic. IEEE Std. 854–1987. New York: IEEE.

Arbel, A. 1993. *Exploring Interior-Point Linear Programming Algorithms and Software*. Cambridge, MA: MIT Press.

Argyros, I. K., and F. Szidarovszky. 1993. *The Theory and Applications of Iteration Methods*. Boca Raton, FL: CRC Press.

Ascher, U. M., R. M. M. Mattheij, and R. D. Russell. 1995. *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*. Philadelphia: SIAM.

Atkinson, K. 1985. *Elementary Numerical Analysis*. New York: Wiley.

Aubin, J. P. 1998. *Optima and Equilibria: An Introduction to Nonliear Analysis*, 2nd ed. New York: Springer-Verlag.

Axelsson, O. 1980. A generalized conjugate direction method and its application on a singular perturbation problem. In *Numerical Analysis*: *LNM* **773**. New York: Springer-Verlag.

Axelsson, O. 1994. *Iterative Solution Methods*. New York: Cambridge University Press.

Ayoub, R. 1974. Euler and the zeta function. *AMM* **81**, 1067–1086.

Aziz, A. K. (ed.). 1969. *Numerical Solution of Differential Equations*. New York: van Nostrand.

Aziz, A. K. (ed.). 1974. *Numerical Solutions of Boundary Value Problems for Ordinary Differential Equations*. New York: Academic Press.

Babuška, I., M. Prager, and E. Vitasék. 1966. *Numerical Processes in Differential Equations*. New York: Wiley-Interscience.

Backus, J. 1979. The history of Fortran I, II, and III. *Annals of the History of Computing* **1**, 21–37.

Bailey, P. B., L. F. Shampine, and P. E. Waltman. 1968. *Nonlinear Two-Point Boundary-Value Problems*. New York: Academic Press.

Bak, J., and D. J. Newman. 1982. *Complex Analysis*. New York: Springer-Verlag.

Baker, C. T. A., C. A. H. Paul, and D. R. Willé. 1995. Issues in the numerical solution of evolutionary delay differential equations. *Advances in Computational Mathematics* **3**, 171–196.

Barnes, E. R. 1986. A variation on Karmarkar algorithm for solving linear programming problems. *MP* **36**, 174–182.

Barnhill, R., R. P. Dube, and F. F. Little. 1983. Properties of Shepard's surfaces. *Rocky Mtn. J. Math.* **13**, 365–382.

Barnhill, R., and A. Riesenfeld. 1974. *Computer Aided Geometric Design*. New York: Academic Press.

Barnsley, M. 1988. *Fractals Everywhere*. New York: Academic Press.

Barnsley, M., and A. Sloan. 1988. A better way to compress images. *Byte* **13**, 215–223.

Barrodale, I., and C. Phillips. 1975. Solution of an overdetermined system of linear equations in the Chebyshev norm. *ACM-TOMS* **1**, 264–270.

Barrodale, I., and F. D. K. Roberts. 1974. Solution of an overdetermined system of equations in the ℓ_1 norm. *ACM-COM* **17**, 319–320.

Barrodale, I., F. D. K. Roberts, and B. L. Ehle. 1971. *Elementary Computer Applications*. New York: Wiley.

Bartels, R. H. 1971. A stabilization of the simplex method. *NM* **16**, 414–434.

Bartels, R., J. Beatty, and B. Barsky. 1987. *An Introduction to Splines for Use in Computer Graphics and Geometric Modeling*. Los Altos, CA: Morgan Kaufmann.

Bartle, R. G. 1976. *The Elements of Real Analysis*. 2nd ed. New York: Wiley.

Becker, E. B., G. F. Carey, and J. T. Oden. 1981. *Finite Elements: An Introduction*. Vol. 1. Englewood Cliffs, NJ: Prentice-Hall.

Bell, E. T. 1975. *Men of Mathematics*. New York: Simon & Schuster.

Bell, G., and S. Glasstone. 1970. *Nuclear Reactor Theory*. New York: van Nostrand-Reinhold.

Bellman, R., and K. L. Cooke. 1963. *Differential-Difference Equations*. New York: Academic Press.

Belsley, D. A., E. Kuh, and R. Welsch. 1981. *Regression Diagnostics: Identifying Influential Data and Sources of Colinearity*. New York: Wiley.

Bender, C. M., and S. A. Orszag. 1978. *Advanced Mathematical Methods for Scientists and Engineers*. New York: McGraw-Hill.

Bhatti, M.A., 2000. *Practical Optimization Methods*. New York: Springer-Verlag.

Birkhoff, G., and R. E. Lynch. 1984. *Numerical Solution of Elliptic Problems*. Philadelphia: SIAM.

Bischof, C., A. Carle, P. Khademi, and A. Mauer. 1994. The ADIFOR 2.0 system for the automatic differentiation of Fortran 77 programs. Mathematics and Computer Sciences Report ANL/MCS-P481-1194. Argonne, IL: Argonne National Laboratory.

Björck, Å. 1967. Solving linear least squares problems by Gram-Schmidt orthogonalization. *BIT* **7**, 1–21.

Björck, Å., and C. C. Paige. 1992. Loss and recapture of orthogonality in the modified Gram-Schmidt algorithm. *SIAM-MAA* **13**, 176–190.

Bloomfield, P. 1976. *Fourier Analysis of Time Series: An Introduction*. New York: Wiley-Interscience.

Blum, E. K. 1972. *Numerical Analysis and Computation: Theory and Practice*. Reading, MA: Addison-Wesley.

Bodewig, E. 1946. Sur la méthod de Laguerre pour l'approximation des racines de certaines équations algébriques et sur la critique d'Hermite. *Nederl. Acad. Wetensch. Proc.* **49**, 911–921.

Boggs, P., R. H. Byrd, and R. B. Schnabel. 1985. *Numerical Optimization 1984*. Philadelphia: SIAM.

Bohman, H. 1952. On approximation of continuous and analytic functions. *Arkiv för Matematik* **2**, 43–56.

Boisvert, R. F., S. E. Howe, D. K. Kahaner, and J. L. Springmann. 1990. Guide to available mathematical software. Center for Computing and Applied Mathematics. Gaithersburg, MD: National Institute of Standards and Technology.

Boisvert, R. F., and R. A. Sweet. 1982. Sources and development of mathematical software for elliptic boundary value problems. In *Sources and Development of Mathematical Software* (W. Cowell, ed.). Englewood Cliffs, NJ: Prentice-Hall.

Borwein, J., and A.S. Lewis, 2000. *Convex Analysis and Nonlinear Optimization*. New York: Springer-Verlag.

de Boor, C. 1971. CADRE: An algorithm for numerical quadrature. In *Mathematical Software* (J. R. Rice, ed.). New York: Academic Press.

de Boor, C. 1976. Total positivity of the spline collocation matrix. *IU-JM* **25**, 541–551.

de Boor, C. 1984. *A Practical Guide to Splines*. 2nd ed. New York: Springer-Verlag.

de Boor, C., and G. H. Golub (eds.). 1978. *Recent Advances in Numerical Analysis*. New York: Academic Press.

Borwein, J. M., and P. B. Borwein. 1984. The arithmetic-geometric mean and fast computation of elementary functions. *SIAM-REV* **26**, 351–366.

Botha, J. F., and G. F. Pinder. 1983. *Fundamental Concepts in the Numerical Solution of Differential Equations*. New York: Wiley.

Boyce, W. E., and R. C. DiPrima. 1977. *Elementary Differential Equations and Boundary Value Problems*. New York: Wiley.

Braess, D. 1984. *Nonlinear Approximation Theory*. New York: Springer-Verlag.

Bramble, J. H. (ed.). 1966. *Numerical Solution of Partial Differential Equations*. New York: Academic Press.

Bratley, P., B. L. Fox, and L. Schrage. 1987. *A Guide to Simulation*. New York: Springer-Verlag.

Brenan, K. E., S. L. Campbell, and L. R. Petzold. 1995. *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*. Philadelphia: SIAM.

Brent, R. P. 1973. *Algorithms for Minimization without Derivatives*. Englewood Cliffs, NJ: Prentice-Hall.

Brent, R. P. 1976. Fast multiple precision evaluation of elementary functions. *ACM-J* **23**, 242–251.

Brezinski, C. 1994. The generalizations of Newton's interpolation formula due to Mühlbach and Andoyer. *ETNA* **2**, 130–137.

Briggs, W. T. 1987. *A Multigrid Tutorial*. Philadelphia: SIAM.

Briggs, W. T., and V. E. Henson. 1995. *The DFT: An Owner's Manual for the Discrete Fourier Transform*. Philadelphia: SIAM.

Brigham, E. O. 1974. *The Fast Fourier Transform*. Englewood Cliffs, NJ: Prentice-Hall.

Brophy, J. F., and P. W. Smith. 1988. Prototyping Karmarkar's algorithm using MATH / PROTAN. *Directions* **5**, 2–3, Houston: Visual Numerics, Inc.

Brown, P. J. (ed.). 1977. *Software Portability*. New York: Cambridge University Press.

Brown, P. N., G. D. Byrne, and A. C. Hindmarsh. 1989. VODE: a variable coefficient ODE solver. *SIAM-SSC* **10**, 1039–1051.

Buchanan, J. L., and P.R. Turner. 1992. *Numerical Methods and Analysis*. New York: McGraw-Hill.

Bunch, J. R., and D. J. Rose (eds.). 1976. *Sparse Matrix Computations*. New York: Academic Press.

Burden, R. L., and J. D. Faires. 1993. *Numerical Analysis*. 5th ed. Boston: PWS-Kent.

Burrage, K. 1978. A special family of Runge-Kutta methods for solving stiff differential equations. *BIT* **18**, 22–41.

Burrage, K. 1995. *Parallel and Sequential Methods for Ordinary Differential Equations*. New York: Oxford University Press.

Butcher, J. C. 1987. *The Numerical Analysis of Ordinary Differential Equations: Runge-Kutta and General Linear Methods*. New York: Wiley.

Buzbee, B. L. 1984. The SLATEC common mathematical library. In *Sources and Development of Mathematical Software* (W. R. Cowell, ed.). Englewood Cliffs, NJ: Prentice-Hall.

Byrne, G. D., and C. A. Hall (eds.). 1973. *Numerical Solution of Systems of Nonlinear Algebraic Equations*. New York: Academic Press.

Byrne, G., and A. Hindmarsh. 1987. Stiff ODE solvers: A review of current and coming attractions. *JCP* **70**, 1–62.

Calvo, M., J. I. Montijano, and L. Rández. 1993. On the change of stepsizes in multistep codes. *Num. Alg.* **4**, 283–304.

Carter, L. L., and E. D. Cashwell. 1975. Particle-transport with the Monte Carlo method. *ERDA Critical Review Series* TID–26607. Springfield, VA: National Technical Information Service.

Cash, J. R. 1979. *Stable Recursions*. New York: Academic Press.

Cassels, J. W. S. 1981. *Economics for Mathematicians*. New York: Cambridge University Press.

Chaitlin, G. J. 1975. Randomness and mathematical proof. *SA* May, 47–52.

Chambers, J. M. 1977. *Computational Methods for Data Analysis*. New York: Wiley.

Chatterjee, S., and B. Price. 1977. *Regression Analysis by Example*. New York: Wiley.

Cheney, E. W. 1982. *Introduction to Approximation Theory*. New York: Chelsea.

Cheney, W. 2001. *Analysis for Applied Mathematics*. New York: Springer-Verlag.

Cheney, W., and D. Kincaid. 1999. *Numerical Mathematics and Computing*. 4th ed. Pacific Grove, CA: Brooks/Cole.

Cheney, W., and W. Light, 1999. *A Course in Approximation Theory*. Pacific Grove, CA: Brooks/Cole.

Cherkasova, M. P. 1972. *Collected Problems in Numerical Analysis*. Groningen, Netherlands: Wolters-Noordhoff.

Childs, B., M. Scott, J. W. Daniel, E. Denman, and P. Nelson (eds.). 1979. *Codes for Boundary Value Problems in Ordinary Differential Equations: LNCS* **76**. New York: Springer-Verlag.

Chow, S. N., J. Mallet-Paret, and J. A. Yorke. 1978. Finding zeros of maps: Homotopy methods that are constructive with probability one. *MOC* **32**, 887–899.

Chui, C. K. 1988. Multivariate splines. *SIAM Regional Conference Series in Mathematics*, Vol. 54.

Chung, K. C., and T. H. Yao. 1977. On lattices admitting unique Lagrange interpolations. *SIAM-NA* **14**, 735–743.

Cline, A. K. 1974a. Scalar and planar valued curve-fitting using splines under tension. *ACM-COM* **17**, 218–220.

Cline, A. K. 1974b. Six subprograms for curve-fitting using splines under tension. *ACM-COM* **17**, 220–223.

Cline, A. K., C. B. Moler, G. W. Stewart, and J. H. Wilkinson. 1979. An estimate for the condition number of a matrix. *SIAM-NA* **16**, 368–375.

Coddington, E. A., and N. Levinson. 1955. *Theory of Ordinary Differential Equations*. New York: McGraw-Hill.

Cody, W. J. 1988. Floating-point standards—theory and practice. In *Reliability in Computing*, 99–107. New York: Academic Press.

Cody, W. J., and W. Waite. 1980. *Software Manual for the Elementary Functions*. Englewood Cliffs, NJ: Prentice-Hall.

Cohen, A. M. 1974. A note on pivot size in Gaussian elimination. *LAA* **8**, 361–368.

Coleman, T. F., and C. van Loan. 1988. *Handbook for Matrix Computations*. Philadelphia: SIAM.

Collatz, L. 1966a. *Functional Analysis and Numerical Mathematics*. 3rd ed. New York: Academic Press.

Collatz, L. 1966b. *The Numerical Treatment of Differential Equations*. New York: Springer-Verlag.

Concus, P., G. H. Golub, and D. P. O'Leary. 1976. A generalized conjugate gradient method for the numerical solution of elliptical partial differential equations. In *Sparse Matrix Computations* (J. R. Bunch and D. J. Rose, eds.). New York: Academic Press.

Conn, A. R., N. I. M. Gould, and Ph. L. Toint. 2000. *Trust-Region Methods*. Philadelphia: SIAM.

Conte, S. D., and C. de Boor. 1980. *Elementary Numerical Analysis*. 3rd ed. New York: McGraw-Hill.

Cooley, J. W., P. A. Lewis, and P. P. Welch. 1967. Historical notes on the fast Fourier transform. *Proceedings IEEE* **55**, 1675–1677.

Coonen, J. T. 1980. An implementation guide to a proposed standard for floating-point arithmetic. *Computer* **13**, 68–79.

Coonen, J. T. 1981. Underflow and the denormalized numbers. *Computer* **14**, 75–87.

Cornuejols, G. et al. (eds.). 2000. *Integer Programming and Combinatorial Optimization.* New York: Springer-Verlag.

Cowell, W. (ed.). 1977. Portability of numerical software. In *LNCS* **57**. New York: Springer-Verlag.

Crowder, H., R. S. Dembo, and J. M. Mulvey. 1979. On reporting computational experiments with mathematical software. *ACM-TOMS* **5**, 193–203.

Cryer, C. W. 1968. Pivot size in Gaussian elimination. *NM* **12**, 335–345.

Cullum, J., and R. A. Willoughby (eds.). 1986. *Large Scale Eigenvalue Problems.* Amsterdam: Elsevier.

Curry, J. H., L. Garnett, and D. Sullivan. 1983. On the iteration of a rational function: Computer experiments with Newton's method. *CMP* **91**, 267–277.

Dahlquist, G. 1956. Convergence and stability in the numerical integration of ordinary differential equations. *Math. Scand.* **4**, 33–35.

Dahlquist, G. 1963. A special stability problem for linear multistep methods. *BIT* **3**, 27–43.

Dahlquist, G., and A. Björck. 1974. *Numerical Methods.* Englewood Cliffs, NJ: Prentice-Hall.

Daniel, J. W., and R. E. Moore. 1970. *Computation and Theory in Ordinary Differential Equations.* San Francisco: Freeman.

Daniels, R.W. 1978. *An Introduction to Numerical Methods and Optimization Techniques.* New York: North-Holland.

Dano, S. 1974. *Linear Programming in Industry.* 4th ed. New York: Springer-Verlag.

Dantzig, G. B. 1948. Programming in a linear structure. Washington, DC: U.S. Air Force, Comptroller's Office.

Dantzig, G. B. 1963. *Linear Programming and Extensions.* Princeton, NJ: Princeton University Press.

Datta, B. N. 1995. *Numerical Linear Algebra and Applications* Pacific Grove, CA: Brooks/Cole.

Davidon, W. C. 1959. Variable metric method for minimization, Research and Development Report ANL-5990 (Rev.) Atomic Energy Commission.

Davis, H. T. 1962. *Introduction to Nonlinear Differential and Integral Equations.* New York: Dover.

Davis, P. J. 1982. *Interpolation and Approximation.* New York: Dover.

Davis, P. J., and P. Rabinowitz. 1956. Abscissas and weights for Gaussian quadratures of high order. *JR-NBS* **56**, 35–37.

Davis, P. J., and P. Rabinowitz. 1984. *Methods of Numerical Integration.* 2nd ed. New York: Academic Press.

Day, J., and B. Peterson. 1988. Growth in Gaussian elimination. *AMM* **95**, 489–513.

Dejon, B., and P. Henrici (eds.). 1969. *Constructive Aspects of the Fundamental Theory of Algebra.* New York: Wiley.

Dekker, K., and J. G. Verwer. 1984. *Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations.* Amsterdam: Elsevier Science.

Dekker, T. J. 1969. Finding a zero by means of successive linear interpolation. In *Constructive Aspects of the Fundamental Theorem of Algebra* (B. Dejon and P. Henrici, eds.). New York: Wiley-Interscience.

Delves, L. M., and J. Mohamed. 1985. *Computational Methods for Integral Equations.* New York: Cambridge University Press.

Demmel, J., and K. Veselić. 1992. Jacobi's method is more accurate than QR. *SIAM-MAA* **13**, 1204–1245.

Dennis, J. E., Jr., and J. Moré. 1974. Quasi-Newton methods, motivation and theory. *SIAM-REV* **19**, 46–89.

Dennis, J. E., Jr., and R. B. Schnabel. 1983. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Englewood Cliffs, NJ: Prentice-Hall.

Dennis, J. E., Jr., and D. J. Woods. 1987. Optimization on microcomputers: The Nelder-Meade simplex algorithm, in *New Computing Environments*, A. Wouk (ed.). Philadelphia: SIAM.

Deuflhard, P., and G. Heindl. 1979. Affine invariant convergence theorems for Newton's method and extensions to related methods. *SIAM-NA* **16**, 1–10.

Dewdney, A. K. 1988. Computer recreations: Random walks that lead to fractal crowds. *SA*, December.

Diekmann, O., S. A. van Gils, S. M. Verduyn Lunel, and H. O. Walther. 1995. Delay equations. New York: Springer-Verlag.

Dieudonné, J. 1960. *Foundations of Modern Analysis*. New York: Academic Press.

de Doncker, E., and I. Robinson. 1984. An algorithm for automatic integration over a triangle using nonlinear extrapolation. *ACM-TOMS* **10**, 1–16.

Dongarra, J. J., J. R. Bunch, C. B. Moler, and G. W. Stewart. 1979. *Users Guide*. Philadelphia: SIAM.

Dongarra, J. J., and D. W. Walker. 1995. Software libraries for linear algebra computations on high performance computers. *SIAM-REV* **37**, 151–180.

Draper, N. R., and H. Smith. 1981. *Applied Regression Analysis*. New York: Wiley.

Driver, R. 1977. *Ordinary and Delay Differential Equations*. New York: Springer-Verlag.

Duff, I. S., A. M. Erisman, and J. K. Reid. 1986. *Direct Methods for Sparse Matrices*. New York: Oxford University Press.

Duffy, D. G. 1993. On the numerical inversion of Laplace transforms: Comparison of three new methods on characteristic problems from applications. *ACM-TOMS* **19**, 333–359.

Durand, E. 1960. Solutions Numériques des Équations Algébriques. (2 vols.) Paris: Mason.

Eaves, B. C. 1976. A short course in solving equations with PL homotopies. *SIAM-AMS Proceedings* **9**, 73–144.

Eaves, B. C., F. J. Gould, H.-O. Peitgen, and M. J. Todd (eds.). 1983. *Homotopy Methods and Global Convergence*. New York: Plenum.

Edelman, A. 1992. The complete pivoting conjecture for Gaussian elimination is false. Deptment of Mathematics. Berkeley, CA: Lawrence Berkeley National Laboratory and University of California, Berkeley.

Edelman, A. 1994. When is $x * (1/x) \neq 1$? Department of Mathematics. Cambridge, MA: MIT.

Eggermont, P. P. B. 1988. Noncentral difference quotients and the derivative. *AMM* **95**, 551–553.

Elliott, D. F., and K. R. Rao. 1982. *Fast Transforms: Algorithms, Analyses, Applications*. New York: Academic Press.

Engels, H. 1980. *Numerical Quadrature and Cubature*. New York: Academic Press.

Epperson, J. F. 1987. On the Runge example. *AMM* **4**, 329–341.

Farwig, R. 1986. Rate of convergence of Shepard's global interpolation formula. *MOC* **46**, 577–590.

Fatunla, S. O. 1988. *Numerical Methods for Initial Value Problems in Ordinary Differential Equations*. New York: Academic Press.

Fefferman, C. 1967. An easy proof of the fundamental theorem of algebra. *AMM* **74**, 854–855.

Fehlberg, E. 1969. Klassische Runge-Kutta Formeln fünfter und siebenter Ordnung mit Schrittweitenkontrolle. *Computing* **4**, 93–106.

Feldstein, A., and P. Turner. 1986. Overflow, underflow, and severe loss of significance in floating-point addition and subtraction. *IMA-JNA* **6**, 241–251.

Ficken, F. A. 1951. The continuation method for functional equations. *CPAM* **4**, 435–456.

Flehinger, B. J. 1966. On the probability that a random integer has initial digit A. *AMM* **73**, 1056–1061.

Fletcher, R., and M. J. D. Powell. 1963. A rapidly convergent descent method for minimization, *CJ* 6, 163–168.

Forsythe, G. E. 1957. Generation and use of orthogonal polynomials for data-fitting with a digital computer. *SIAM Journal* **5**, 74–88.

Forsythe, G. E., M. A. Malcolm, and C. B. Moler. 1977. *Computer Methods for Mathematical Computations*. Englewood Cliffs, NJ: Prentice-Hall.

Forsythe, G. E., and C. B. Moler. 1967. *Computer Solution of Linear Algebraic Systems*. Englewood Cliffs, NJ: Prentice-Hall.

Forsythe, G. E., and W. R. Wasow. 1960. *Finite-Difference Methods for Partial Differential Equations*. New York: Wiley.

Fosdick, L. D. (ed.). 1979. *Performance Evaluation of Numerical Software*. Amsterdam: North-Holland.

Fosdick, L. D. 1993. IEEE Arithmetic Short Reference. High Performance Scientific Computing, Boulder, CO: University of Colorado at Boulder.

Foster, L. V. 1981. Generalizations of Laguerre's method: Higher order methods. *SIAM-NA* **18**, 1004–1018.

Foster, L. V. 1994. Gaussian elimination with partial pivoting can fail in practice. *SIAM-MAA* **15**, 1354–1362.

Fournier, A., D. Fussell, and L. Carpenter. 1982. Computer rendering of stochastic models. *ACM-COM* **25**, 371–384.

Fox, L. 1987. *Biographical Memoirs of Fellows of the Royal Society: James Hardy Wilkinson 1919–1986*, **33**. London: Royal Society.

Fox, P. A., A. D. Hall, and N. L. Schryer. 1978. Framework for a portable library. *ACM-TOMS* **4**, 177–188.

Francis, J. G. F. 1961. The QR transformation: A unitary analogue to the LR transformation. Parts 1 and 2. *CJ* **4**, 265–272, 332–345.

Franke, R. 1982. Scattered data interpolation: Tests of some methods. *MOC* **38**, 181–200.

Fritsch, F. N., and R. E. Carlson. 1980. Monotone piecewise cubic interpolation. *SIAM-NA* **17**, 238–246.

Fröberg, C. E. 1969. *Introduction to Numerical Analysis*. 2nd ed. Reading, MA: Addison-Wesley.

Gaffney, P. 1987. When things go wrong... Report BSC87/1. Bergen, Norway: IBM Bergen Scientific Centre.

Galeone, L. 1977. Generalizzazione del methodo di Laguerre. *Calcolo* **14**, 121–131.

Garbow, B. S., J. M. Boyle, J. J. Dongarra, and C. B. Moler. 1972. *Matrix Eigensystem Routines: Guide Extension*. New York: Springer-Verlag.

Garcia, C. B., and F. J. Gould. 1980. Relations between several path-following algorithms and local and global Newton methods. *SIAM-REV* **22**, 263–274.

Garcia, C. B., and W. I. Zangwill. 1981. *Pathways to Solutions, Fixed Points, and Equilibria*. Englewood Cliffs, NJ: Prentice-Hall.

Gardner, M. 1961. *Mathematical Puzzles and Diversions*. New York: Simon & Schuster.

Gasca, M., and J. I. Maeztu. 1982. On Lagrange and Hermite interpolation in $\mathbb{R}^k$. *NM* **39**, 1–14.

Gautschi, W. 1961. Recursive computation of certain integrals. *ACM-J* **8**, 21–40.

Gautschi, W. 1967. Computational aspects of three-term recurrence relations. *SIAM-REV* **9**, 24–82.

Gautschi, W. 1975. Computational methods in special functions. In *Theory and Applications of Special Functions* (R. Askey, ed.). New York: Academic Press, 1–98.

Gautschi, W. 1976. Advances in Chebyshev quadrature. In *Numerical Analysis* (G. A. Watson, ed.). *LNM* **506**. New York: Springer-Verlag.

Gautschi, W. 1979. Families of algebraic test equations. *Calcolo* **16**, 383–398.

Gautschi, W. 1983. How and how not to check Gaussian quadrature formulae. *BIT* **23**, 209–216.

Gautschi, W. 1984. Questions of numerical condition related to polynomials. In *Studies in Numerical Analysis* (G. H. Golub, ed.), 140–177. Washington, DC: MAA.

Gear, C. W. 1971. *Numerical Initial Value Problems in Ordinary Differential Equations*. Englewood Cliffs, NJ: Prentice-Hall.

Gekeler, E. 1984. *Discretization Methods for Stable Initial Value Problems*: *LNM* **1044**. New York: Springer-Verlag.

Gentleman, W. M. 1972. Implementing Clenshaw-Curtis quadrature. *ACM-J* **15**, 337–342.

George, A., and J. W. Liu. 1981. *Computer Solution of Large Sparse Positive Definite Systems*. Englewood Cliffs, NJ: Prentice-Hall.

George, A., J. W. Liu, and E. Ng. 1980. User guide for SPARSPACK: Waterloo sparse linear equations package. Computer Science Department Report CS-78-30 (revised 1980). Waterloo, Canada: University of Waterloo.

Gerald, C. F., and P. O. Wheatley. 1989. *Applied Numerical Analysis*. 4th ed. Reading, MA: Addison-Wesley.

Ghizetti, A., and A. Ossiccini. 1970. *Quadrature Formulæ*. New York: Academic Press.

Gill, P. E., G. H. Golub, W. Murray, and M. A. Saunders. 1974. Methods for modifying matrix factorizations. *MOC* **28**, 505–535.

Gill, P. E., and W. Murray. 1974. Newton-type methods for unconstrained and linearly constrained optimization. *MP* **28**, 311–350.

Gill, P. E., W. Murray, and M. H. Wright. 1981. *Practical Optimization*. New York: Academic Press.

Gladwell, I., L. F. Shampine, and R. W. Brankin. 1987. Automatic selection of the initial stepsize for an ODE solver. *JCAM* **18**, 175–192.

Gladwell, J., and R. Wait. 1979. *A Survey of Numerical Methods for Partial Differential Equations*. New York: Oxford University Press.

Glatz, G. 1978. Stabile Deflationsalgorithmen bei der numerischen Berechnung von Polynomnullstellen. *ZAMM* **58**, T416–T418.

Glieck, J. 1987. *Chaos*. New York: Viking Press.

Goldstein, A. A. 1967. *Constructive Real Analysis*. New York: Harper & Row.

Goldstine, H. H. 1977. *A History of Numerical Analysis from the 16th Through the 19th Century*. New York: Springer-Verlag.

Golub, G. H. (ed.). 1984. *Studies in Numerical Analysis*. Washington, DC: MAA.

Golub, G. H., and D. P. O'Leary. 1989. Some history of the conjugate gradient and Lanczos methods. *SIAM-REV* **31**, 50–102.

Golub, G. H., and J. M. Ortega. 1992. *Scientific Computing and Differential Equations*. New York: Academic Press.

Golub, G. H., and C. F. van Loan. 1980. An analysis of the total least squares problem. *SIAM-NA* **17**, 883–893.

Golub, G. H., and C. F. van Loan. 1989. *Matrix Computations*. 2nd ed. Baltimore, MD: Johns Hopkins University Press.

Gonzaga, C. C. 1992. Path-following methods for linear programming. *SIAM-REV* **34**, 167–224.

Good, I. J. 1972. What is the most amazing approximate integer in the universe? *Pi Mu Epsilon Journal* **5**, 314–315.

Gordon, W. J., and J. A. Wixom. 1978. Shepard's method of 'metric interpolation' to bivariate and multivariate interpolation. *MOC* **32**, 253–264.

Gould, N. 1991. On growth in Gaussian elimination with complete pivoting. *SIAM-MAA* **12**, 354–361.

Gourlay, A. R., and G. A. Watson. 1973. *Computational Methods for Matrix Eigenvalues*. New York: Wiley.

Greenspan, D. 1965. *Introductory Numerical Analysis of Elliptic Boundary Value Problems*. New York: Harper & Row.

Gregory, J. A. (ed.). 1986. *The Mathematics of Surfaces*. New York: Oxford University Press.

Gregory, R. T. 1980. *Error-Free Computation*. Huntington, NY: Krieger.

Gregory, R. T., and D. Karney. 1969. *A Collection of Matrices for Testing Computational Algorithms*. New York: Wiley.

Griewank, A. 2000. *Evaluating Derivative Principles and Techniques of Algorithmic Differentiation*. Philadelphia: SIAM

Griewank, A., and G. F. Corliss. 1991. *Automatic Differentiation of Algorithms: Theory, Implementation, and Applications*. Philadelphia: SIAM.

Griffiths, P., and J. Harris. 1978. *Principles of Algebraic Geometry*. New York: Wiley.

Gustafson, B., and J. Oliger. 1995. *Time Dependent Problems and Difference Equations*. New York: Wiley.

Haar, A. 1918. Die minkowskische Geometrie und die Annäherung an stetige Funktionen. *Mathematische Annalen* **78**, 294–311.

Haber, S. 1970. Numerical Evaluation of Multiple Integrals. *SIAM-REV* **12**, 481–526.

Haberman, R. 1977. *Mathematical Models*. Englewood Cliffs, NJ: Prentice-Hall.

Hackbusch, W. 1995. *Iterative Solution of Large Sparse Systems of Equations*. New York: Springer-Verlag.

Hackbusch, W., and U. Trottenberg (eds.). 1982. *Multigrid Methods: LNM* **960**. New York: Springer-Verlag.

Hageman, L. A., and D. M. Young. 1981. *Applied Iterative Methods*. New York: Academic Press.

Hairer, E., S. P. Nörsett, and G. Wanner. 1987. *Solving Ordinary Differential Equations I—Nonstiff Problems*. New York: Springer-Verlag.

Hairer, E., S. P. Nörsett, and G. Wanner. 1991. *Solving Ordinary Differential Equations II—Stiff and Differential-Algebraic Problems*. New York: Springer-Verlag.

Hammerlin, G. (ed.). 1982. *Numerical Integration*. New York: Birkhäuser-Verlag.

Hammersley, J. M., and DC Handscomb. 1964. *Monte Carlo Methods*. London: Methuen.

Hamming, R. W. 1973. *Numerical Methods for Scientists and Engineers*. New York: McGraw-Hill.

Hansen, E. R. 1969. *Topics in Interval Analysis*. New York: Oxford University Press.

Hardy, G. H. 1960. *A Course of Pure Mathematics*. 10th ed.. New York: Cambridge University Press.

Hardy, R. L. 1971. Multiquadric equations of topography and other irregular surfaces. *Journal Geophysical Research* **76**, 1905–1915.

Hart, J. F., E. W. Cheney, C. L. Lawson, H. J. Maehly, C. K. Mesztenyi, J. R. Rice, H. G. Thacher, Jr., and C. Witzgall. 1968. *Computer Approximations*. New York: Wiley. (Reprinted Huntington, NY: Krieger, 1978.)

Hartley, P. H. 1976. Tensor product approximations to data defined on rectangle meshes in n-space. *CJ* **19**, 348–352.

Heller, D. 1978. A survey of parallel algorithms in numerical linear algebra. *SIAM-REV* **20**, 740–777.

Hennell, M. A., and L. M. Delves (eds.). 1980. *Production and Assessment of Numerical Software*. New York: Academic Press.

Henrici, P. 1962. *Discrete Variable Methods in Ordinary Differential Equations*. New York: Wiley.

Henrici, P. 1963. *Error Propagation for Difference Methods*. New York: Wiley.

Henrici, P. 1964. *Elements of Numerical Analysis*. New York: Wiley.

Henrici, P. 1974. *Applied and Computational Complex Analysis* (3 volumes). New York: Wiley.

Hestenes, M. R. 1980. *Conjugate Direction Methods in Optimization*. New York: Springer-Verlag.

Hestenes, M. R., and E. Stiefel. 1952. Methods of conjugate gradient for solving linear systems. *JR-NBS* **45**, 409–436.

Hestenes, M. R., and J. Todd. 1991. *Mathematicians Learning to Use Computers*. Special Publication 730. Gaithersburg, MD: National Institute of Standards and Technology.

Hetzel, W. C. (ed.). 1973. *Program Test Methods*. Englewood Cliffs, NJ: Prentice-Hall.

Higham, N. J. 1996. *Accuracy and Stability of Numerical Algorithms*. Philadelphia: SIAM

Higham, N. J., and D. J. Higham. 1989. Large growth factors in Gaussian elimination with pivoting. *SIAM-MAA* **10**, 155–164.

Higham, N. J., and N. Trefethen. 1991. Complete pivoting conjecture is disproved. *SIAM News* **24**, 9.

Hindmarsh, A. 1980. LSODE and LSODEI: Two initial value ordinary differential equations solvers. *ACM Special Interest Group in Numerical Methods Newsletter* **15**, 10–11.

Hirsch, M. W., and S. Smale. 1979. On algorithms for solving $f(x) = 0$. *CPAM* **32**, 281–312.

Holland. J. H. 1989. Searching nonlinear functions for high values, *AMC* **32** 255–274.

Holmes, R. B. 1972. *A Course on Optimization and Best Approximation*. New York: Springer-Verlag.

Horn, R. A., and C. R. Johnson. 1986. *Matrix Analysis*. New York: Cambridge University Press.

Hough, D. 1981. Applications of the proposed IEEE 754 standard for floating-point arithmetic. *Computer* **14**, 70–74.

Householder, A. S. 1964. *The Theory of Matrices in Numerical Analysis*. New York: Blaisdell. (Reprinted New York: Dover, 1974.)

Householder, A. S. 1970. *The Numerical Treatment of a Single Nonlinear Equation*. New York: McGraw-Hill.

Hull, T. E., W. H. Enright, B. M. Fellen, and A. E. Sedgwick. 1972. Comparing numerical methods for ordinary differential equations. *SIAM-NA* **9**, 603–637.

IEEE. 1981. A proposed standard for binary floating-point arithmetic: Draft 8.0 of IEEE Task P754. *Computer* **14**, March.

IEEE. 1985. IEEE standard for binary floating point arithmetic. *ANSI/IEEE Standard* **P754**. New York: IEEE.

IEEE. 1987. A radix-independent standard for floating-point arithmetic. IEEE Std. 754–1987. New York: IEEE.

IMSL. 1995. *International Mathematical and Statistical Libraries Reference Manual*. Houston: Visual Numerics, Inc.

Isaacson, E., and H. B. Keller. 1966. *Analysis of Numerical Methods*. New York: Wiley.

Iserles, A. 1994. Numerical analysis of delay differential equations with variable delays, *Annals of Numer. Math.* **1**, 133–152.

Jackson, K. R., W. H. Enright, and T. E. Hull. 1978. A theoretical criterion for comparing Runge-Kutta formulas. *SIAM-NA* **15**, 618–641.

Jacobs, D. (ed.). 1978. *Numerical Software—Needs and Availability*. New York: Academic Press.

Jain, M. K. 1984. *Numerical Solution of Differential Equations*. 2nd ed. New York: Wiley.

Jenkins, M. A., and J. F. Traub. 1970a. A three-stage algorithm for real polynomials using quadratic iteration. *SIAM-NA* **7**, 545–566.

Jenkins, M. A., and J. F. Traub. 1970b. A three-stage variable-shift iteration for polynomial zeros. *NM* **14**, 252–263.

Jennings, A. 1977. *Matrix Computations for Engineers and Scientists*. New York: Wiley.

Jerome, J. W. 1985. Approximate Newton methods and homotopy for stationary operator equations. *CA* **1**, 271–285.

Johnson, L. W., and R. D. Riess. 1982. *Numerical Analysis*. 2nd ed. Reading, MA: Addison-Wesley.

Joubert, W. D., G. F. Carey, N. A. Berner, A. Kalhan, H. Kohli, A. Lorber, R. T. Mclay, and Y. Shen. 1995. PCG Reference Manual. Center for Numerical Analysis Report CNA–274. Austin, TX: The University of Texas at Austin.

Joyce, D. 1971. Survey of extrapolation processes in numerical analysis. *SIAM-REV* **13**, 435–490.

Kahan, W. 1967. Laguerre's method and a circle which contains at least one zero of a polynomial. *SIAM-NA* **4**, 474–482.

Kahan, W. 1993. A fear of constants and disdain for singularities. Berkeley, CA: University of California, Berkeley.

Kahaner, D. 1970. Matrix description of the fast Fourier transform. *IEEE-TAE* **AU-18**, 422–450.

Kahaner, D. 1978. The fast Fourier transform by polynomial evaluation. *ZAMP* **29**, 387–394.

Kahaner, D., C. Moler, and S. Nash. 1989. *Numerical Methods and Software*. Englewood Cliffs, NJ: Prentice-Hall.

Kaps, P., and P. Rentrop. 1979. Generalized Runge-Kutta methods of order four with step size control for stiff ordinary differential equations. *NM* **33**, 55–68.

Karloy, F. P. 1993. Genetic algorithms for the traveling salesman problem, in *Biological Cybernetics*, 539–546, Berlin: Springer-Verlag

Karmarkar, N. 1984. A new polynomial-time algorithm for linear programming. *Combinatorica* **4**, 373–395.

Karon, J. M. 1978. Computing improved Chebyshev approximations by the continuation method. *SIAM-NA* **15**, 1269–1288.

Kearfott, R. B., M. Dawande, K. Du, and C. Hu. 1994. A portable Fortran 77 interval standard function library. *ACM-TOMS* **20**, 447–459.

Keller, H. B. 1968. *Numerical Methods for Two-Point Boundary-Value Problems*. Waltham, MA: Blaisdel.

Keller, H. B. 1976. *Numerical Solution of Two-Point Boundary Value Problems*. Philadelphia: SIAM.

Keller, H. B. 1977. Numerical solution of bifurcation and nonlinear eigenvalue problems. In *Applications of Bifurcation Theory* (P. Rabinowitz, ed.), 359–384. New York: Academic Press.

Keller, H. B. 1978. Global homotopies and Newton methods. In *Recent Advances in Numerical Analysis* (C. de Boor and G. H. Golub, eds.), 73–94. New York: Academic Press.

Kelley, C. T. 1995. *Iterative Methods for Linear and Nonlinear Equations*. Philadelphia: SIAM.

Kelley, C. T. 1999. *Iterative Methods for Optimization*. Philadelphia: SIAM.

Kennedy, W. J., and J. E. Gentle. 1988. *Statistical Computing*. New York: Dekker.

Kernighan, B. W., and P. J. Plauger. 1974. *The Elements of Programming Style*. New York: McGraw-Hill.

Khovanskii, A. N. 1963. *The Application of Continued Fractions and Their Generalizations to Problems in Approximation Theory*. Groningen, Netherlands: Wolters-Noordhoff.

Kincaid, D. R., and T. C. Oppe. 1988. A parallel algorithm for the general LU factorization. *CANM* **4**, 349–359.

Kincaid, D. R., T. C. Oppe, and D. M. Young. 1989. ITPACKV 2D user's guide. Center for Numerical Analysis Report CNA–232. Austin, TX: The University of Texas at Austin.

Kincaid, D. R., J. R. Respess, D. M. Young, and R. G. Grimes. 1982. ITPACK 2C: A Fortran package for solving large sparse linear systems by adaptive accelerated iterative methods. *ACM-TOMS* **8**, 302–322.

Kincaid, D. R., and D. M. Young. 1979. Survey of iterative methods. In *Encyclopedia of Computer Science and Technology* (J. Belzer, A. G. Holzman, and A. Kent, eds.), 354–391. New York: Dekker.

Kirkpatrick, S., et al. 1983. Optimization by simulated annealing, *Science* **220** 671–680.

Kline, M. 1972. *Mathematical Thought from Ancient to Modern Times*. New York: Oxford University Press.

Knuth, D. E. 1969. *The Art of Computer Programming: Seminumerical Algorithms*. Vol. 2. Reading, MA: Addison-Wesley.

Knuth, D. E. 1979. Mathematical typography. *AMS-B* **2**, 337–372.

Krogh, F. 1970. VODQ/SVDQ/DVDQ: Variable order integrators for the numerical solution of ordinary differential equations. Jet Propulsion Laboratory Technical Brief NPO–11643. Pasadena, CA: California Institute of Technology.

Krylov, V. I. 1962. *Approximate Calculation of Integrals* (Transl.: A. Stroud). New York: Macmillan.

Kuang, Y. 1993. *Delay Differential Equations*. New York: Academic Press.

Kulisch, U., and W. Miranker. 1981. *Computer Arithmetic in Theory and Practice*. New York: Academic Press.

Lakshmikantham, V., and D. Trigiante. 1988. *Theory of Difference Equations, Numerical Methods and Examples*. New York: Academic Press.

Lambert, J. 1973. *Computational Methods in Ordinary Differential Equations*. New York: Wiley.

Lancaster, P. 1966. Error analysis for the Newton-Raphson method. *NM* **9**, 55–68.

Lancaster, P., and K. Salkauskas. 1986. *Curve and Surface Fitting*. New York: Academic Press.

Lancaster, P., and M. Tismenetsky. 1985. *Theory of Matrices*. 2nd ed. New York: Academic Press.

Lanczos, C. 1966. *Discourse on Fourier Series*. Edinburgh: Oliver and Boyd.

Lapidus, L., and W. E. Schiesser. 1976. *Numerical Methods for Differential Equations*. New York: Academic Press.

Lapidus, L., and J. Seinfield. 1971. *Numerical Solution of Ordinary Differential Equations*. New York: Academic Press.

Lau, H. T. 1994. *A Numerical Library in C for Scientists and Engineers*. Boca Raton, FL: CRC Press.

Laurie, D. 1978. Automatic numerical integration over a triangle. Technical Report. Pretoria, South Africa: National Research Center for Mathematical Sciences.

Lawrence, D. 1991. *Handbook of Genetic Algorithms*. New York: Reinhold.

Lawson, C. L., and R. J. Hanson. 1995. *Solving Least Squares Problems*. Philadelphia: SIAM.

Lawson, C. L., R. J. Hanson, D. R. Kincaid, and F. T. Krogh. 1979. Basic linear algebra subprograms for Fortran usage. *ACM-TOMS* **5**, 308–323.

Le, D. 1985. An efficient derivative-free method for solving nonlinear equations. *ACM-TOMS* **11**, 250–262.

Lee, S. L. 1994. A note on the total least squares problem for coplanar points. Mathematical Sciences Section Report ORNL/TM-12852. Oak Ridge, TN: Oak Ridge National Laboratory.

Levin, M. 1982. An iterative method for the solution of systems of nonlinear equations. *Analysis* **2**, 305–313.

Li, T. -Y. 1987. Solving polynomial systems. *MI* **9**, 33–39.

Linear, P. 1979. *Theoretical Numerical Analysis*. New York: Wiley.

Longley, J. W. 1984. *Least Squares Computations Using Orthogonalization Methods*. New York: Dekker.

Lorentz, G. G., K. Jetter, and S. D. Riemenschneider. 1983. *Birkhoff Interpolation*. Reading, MA: Addison-Wesley.

Lozier, D. W., and F. W. J. Olver. 1994. Numerical evaluation of special functions. In *Mathematics of Computation 1943–1993: A Half-Century of Computational Mathematics* **48**, 79–125. Providence, RI : AMS.

Luenberger, D. G. 1973. *Introduction to Linear and Nonlinear Programming*. Reading, MA: Addison-Wesley.

Lyness, J. 1983. AUG2: Integration over a triangle. Mathematics and Computer Sciences Report ANL/MCS-TM-13. Argonne, IL: Argonne National Laboratory.

Lyness, J. N., and J. J. Kaganove. 1976. Comments on the nature of automatic quadrature routines. *ACM-TOMS* **2**, 65–81.

Machura, M., and R. Sweet. 1980. Survey of software for partial differential equations. *ACM-TOMS* **6**, 461–488.

MacLeod, M. A. 1973. Improved computation of cubic natural splines with equi-spaced knots. *MOC* **27**, 107–109.

Mandelbrot, B. 1982. *The Fractal Geometry of Nature*. New York: Freeman.

Mangasarian, O. L. 1969. *Computational Methods in Optimization*. New York: McGraw-Hill.

Marchuk, G. I. 1994. *Numerical Methods and Applications*. Boca Raton, FL: CRC Press.

Marden, M. 1949. *The Geometry of the Zeros of a Polynomial in a Complex Variable*. Providence, RI: AMS.

Marden, M. 1966. *Geometry of Polynomials*. Providence, RI: AMS.

Maron, M. J., and R. J. Lopez. 1991. *Numerical Analysis: A Practical Approach*. 3rd ed. Belmont, CA: Wadsworth.

Marsaglia, G. 1968. Random numbers fall mainly in the planes. *Proceedings National Academy Sciences* **61**, 25–28.

Marsden, M. J. 1970. An identity for spline functions with applications to variation-diminishing spline approximation. *JAT* **3**, 7–49.

März, R. 1992. Numerical methods for differential algebraic equations. *AN* 141–198.

McCormick, S. F. 1987. *Multigrid Methods*. Philadelphia: SIAM.

McKeenman, W. M. 1962. Algorithm 145: Adaptive numerical integration by Simpson's rule. *ACM-COM* **5**, 604.

McNamee, J. M. 1985. Numerical differentiation of tabulated functions with automatic choice of step-size. *IJNME* **21**, 1171–1185.

Meinguet, J. 1983. Refined error analysis of Cholesky factorization. *SIAM-NA* **20**, 1243–1250.

Meissner, L. P., and E. I. Organick. 1980. FORTRAN77: *Featuring Structured Programming*. Reading, MA: Addison-Wesley.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. 1953. Equations of state calculation by fast computing machines, *J. Chem. Physics* **21** 1087–1092.

Meyer, G. H. 1968. On solving nonlinear equations with a one-parameter operator embedding. *SIAM-NA* **5**, 739–752.

Micchelli, C. A. 1986a. Algebraic aspects of interpolation. In *Approximation Theory* (C. de Boor, ed.) Proceedings of Symposia in Applied Mathematics **36**, 81–102. Providence, RI: AMS.

Micchelli, C. A. 1986b. Interpolation of scattered data: Distance matrices and conditionally positive definite functions. *CA* **2**, 11–22.

Micchelli, C. A., and T. J. Rivlin (eds.). 1977. *Optimal Estimation in Approximation Theory*. New York: Plenum.

Mickens, R. E. 1987. *Difference Equations*. New York: van Nostrand-Reinhold.

Milne, W. E. 1970. *Numerical Solution of Differential Equations*. New York: Dover.

Miranker, W. L. 1980. *Numerical Methods for Stiff Equations and Singular Perturbation Problems*. Boston: Reidel.

Mitchell, A. 1969. *Computational Methods in Partial Differential Equations*. New York: Wiley.

Mitchell, A. R., and R. Wait. 1977. *The Finite Element Method in Partial Differential Equations*. New York: Wiley.

Moler, C. B., and L. P. Solomon. 1970. Use of splines and numerical integration in geometrical acoustics. *Journal Acoustical Society America* **48**, 739–744.

Moler, C. B., and C. F. van Loan. 1978. Nineteen dubious ways to compute the exponential of a matrix. *SIAM-REV* **20**, 801–836.

Moore, R., 1966. *Interval Analysis*. Englewood Cliffs, NJ: Prentice Hall.

Moore, R. E. 1975. *Mathematical Elements of Scientific Computing*. New York: Holt, Reinhart & Winston.

Moore, R. E. 1979. *Methods and Applications of Interval Analysis*. Philadelphia: SIAM.

Moore, R. E. 1994. Numerical solution of differential equations to prescribed accuracy. *Computers Math. Applic*. **28**, 253–261.

Moré, J. J., B. S. Garbow, and K. E. Hillstrom. 1980. User guide for `MINIPACK-1`. Mathematics and Computer Sciences Report ANL-80-74. Argonne, IL: Argonne National Laboratory.

Morgan, A. 1986. A homotopy for solving polynomial systems. *Applied Mathematics and Computation* **18**, 87–92.

Morgan, A. 1987. *Solving Polynomial Systems Using Continuation for Engineering and Scientific Problems*. Englewood Cliffs, NJ: Prentice-Hall.

Morton, K. W., and D. F. Mayers. 1995. *Numerical Solution of Partial Differential Equations*. Cambridge: Cambridge University Press.

Moses, J. 1971. Symbolic integration: The stormy decade. *ACM-COM* **14**, 548–560.

Murota, K., and M. Iri. 1982. Parameter tuning and repeated application of the IMT type transformation in numerical quadrature. *NM* **38**, 347–363.

Murray, W. (ed.). 1972. *Numerical Methods for Unconstrained Optimization*. New York: Academic Press.

Murtagh, B. A., and M. Saunders. 1978. Large-scale linearly constrained optimization. *MP* **14**, 41–72.

NAG. 1995. *NAG Fortran Library Manual*. Downers Grove, IL: NAG, Inc.

Nazareth, J. L. 1986. Homotopy techniques in linear programming. *Algorithmica* **1**, 529–535.

Nazareth, J. L. 1987. *Computer Solution of Linear Programs*. New York: Oxford University Press.

Nazareth, L., and P. Tseng. 1998. Gilding the lily: A variant of the Nelder-Meade algorithm, preprint.

Nelder, J. A., and R. Mead, 1965. A simplex method for function minimization, *CJ* **7**, 308–313.

Nerinckx, D., and A. Haegemans. 1976. A comparison of nonlinear equation solvers. *JCAM* **2**, 145–148.

Neumaier, A. 1990. *Interval Methods for Systems of Equations*. New York: Cambridge University Press.

Newman, D. J., and T. J. Rivlin. 1983. Optimal universally stable interpolation. *Analysis* **3**, 355–367.

Niederreiter, H. 1978. Quasi-Monte Carlo methods. *AMS-B* **84**, 957–1041.

Nielson, G. M. 1974. Some piecewise polynomial alternatives to splines under tension. In *Computer Aided Geometric Design* (R. E. Barnhill and R. F. Riesenfeld, eds.), 209–235. New York: Academic Press.

Nievergelt, J., J. G. Farrar, and E. M. Reingold. 1974. *Computer Approaches to Mathematical Problems*. Englewood Cliffs, NJ: Prentice-Hall.

Nievergelt, Y. 1991. Numerical linear algebra on the HP–28 or how to lie with supercalculators. *AMM* **98**, 539–544.

Nievergelt, Y. 1994. Total least squares: State-of-the-art regression in numerical analysis. *SIAM-REV* **36**, 258–264.

Noble, B., and J. W. Daniel. 1988. *Applied Linear Algebra*. 3rd ed. Englewood Cliffs, NJ: Prentice-Hall.

Nocedal, J., and S. Wright, 1999. *Numerical Optimization*, New York: Springer-Verlag.

Novak, E., K. Ritter, and H. Wozniakowski. 1995. Average-case optimality of a hybrid secant-bisection method. *MOC* **64**, 1517–1540.

Nussbaumer, H. J. 1982. *Fast Fourier Transform and Convolution Algorithms*. New York: Springer.

Oden, J. T. 1972. *Finite Elements of Nonlinear Continua*. New York: McGraw-Hill.

Oden, J. T., and J. N. Reddy. 1976. *An Introduction to the Mathematical Theory of Finite Elements*. New York: Wiley.

Oppe, T. C., W. D. Joubert, and D. R. Kincaid. 1988. NSPCG user's guide, version 1. 0, package for solving large sparse linear systems by various iterative methods. Center for Numerical Analysis Report CNA–216. Austin, TX: The University of Texas at Austin.

Oppe, T. C., and D. R. Kincaid. 1988. Parallel LU-factorization algorithms for dense matrices. In *Supercomputing* (E. N. Houstis, T. S. Papatheodorou, and C. D. Polychronopoulos, eds.), *LNCS* **297**, 576–594. New York: Springer-Verlag.

Ortega, J. M. 1972. *Numerical Analysis: A Second Course*. New York: Academic Press. (Reprinted Philadelphia: SIAM, 1990.)

Ortega, J. M. 1988. *Introduction to Parallel and Vector Solution of Linear Systems*. New York: Plenum.

Ortega, J. M., and W. C. Poole. 1986. *An Introduction to Numerical Methods for Differential Equations*. New York: Wiley.

Ortega, J. M., and W. C. Rheinboldt. 1970. *Iterative Solution of Nonlinear Equations in Several Variables*. New York: Academic Press.

Osborne, M. R. 1966. On Nordsieck's method for the numerical solution of ordinary differential equations. *BIT* **6**, 51–57.

Ostrowski, A. M. 1966. *Solution of Equations and Systems of Equations*. 2nd ed. New York: Academic Press.

Otten, R. H. J. M., and L. P. P. van Ginneken. 1989. *The Annealing Algorithm*, Dordrecht: Kluwer.

Overton, M. 2001. *Numerical Computing with IEEE Floating Point Arithmetic*. Philadelphia: SIAM.

Paddon, D. J., and H. Holstein. 1985. *Multigrid Methods for Integral and Differential Equations*. New York: Oxford University Press.

Pan, V. 1984. How can we speed up matrix multiplication? *SIAM-REV* **26**, 393–415.

Parlett, B. N. 1964. Laguerre's method applied to the matrix eigenvalue problem. *MOC* **18**, 464–485.

Parlett, B. N. 1981. *The Symmetric Eigenvalue Problem*. Englewood Cliffs, NJ: Prentice-Hall.

Parter, S. 1985. *Large Scale Scientific Computation*. New York: Academic Press.

PCGPAK2. 1990. PCGPAK2 user's guide. New Haven, CT: Scientific Computing Associates, Inc.

Pearson, K. 1901. On lines and planes of closest fit to points in space. *Phil. Mag.* **2**, 559–572.

Peitgen, H., and P. Richter. 1986. *The Beauty of Fractals*. New York: Springer-Verlag.

Peitgen, H.-O., D. Saupe, and F. V. Haeseler. 1984. Cayley's problem and Julia sets. *MI* **6**, 11–20.

Penrose, R. 1955. A generalized inverse for matrices. *Proceedings Cambridge Phil. Society* **51**, 406–413.

Pereyra, V. 1984. Finite difference solution of boundary value problems in ordinary differential equations. In *Studies in Numerical Analysis* (G. H. Golub, ed.), 243–269. Washington, DC: MAA.

Perron, O. 1929. *Die Lehre von Kettenbrüchen*. Leipzig, Germany: Teubner. (Reprinted New York: Chelsea.)

Peters, G., and J. H. Wilkinson. 1971. Practical problems arising in the solution of polynomial equations. *IMA-JNA* **8**, 16–35.

Pham, D. T., and D. Karaboga. 2000. *Intelligent Optimisation Techniques: Genetic Algorithms, Tabu Search, Simulated Annealing and Neural Networks*. New York: Springer-Verlag.

Phillips, G. M., and P. J. Taylor. 1973. *Theory and Applications of Numerical Analysis*. New York: Academic Press.

Pickering, M. 1986. *An Introduction to Fast Fourier Transform Methods for Partial Differential Equations, With Applications*. New York: Wiley.

Pickover, C. A. 1988. A note on chaos and Halley's methods. *ACM-COM* (11) **31**, 11.

Piessens, R., E. deDoncker-Kapenga, C. W. Überhuber, and D. H. Kahaner. 1983. QUADPACK: *A Subroutine Package for Automatic Integration*. New York: Springer-Verlag.

Powell, M. J. D. 1964. An efficient method for finding stationary values of a function of several variables, *CJ* **7**, 155–162.

Powell, M. J. D. (ed.). 1981. *Nonlinear Optimization*, New York: Academic Press.

Powers, D. 1972. *Boundary Value Problems*. New York: Academic Press.

Prager, W. H. 1988. *Applied Numerical Linear Algebra*. Englewood Cliffs, NJ: Prentice-Hall.

Prenter, P. 1975. *Splines and Variational Methods*. New York: Wiley.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. 1986. *Numerical Recipes*. New York: Cambridge University Press.

Prince, P. J., and J. R. Dormand. 1981. High order embedded Runge-Kutta formulæ. *JCAM* **1**, 67–75.

Pritsker, A. 1986. *Introduction to Simulation and SLAM II*. New York: Wiley.

Pruess, S. 1976. Properties of splines in tension. *JAT* **17**, 86–96.

Pruess, S. 1978. An algorithm for computing smoothing splines in tension. *Computing* **19**, 365–373.

Rabinowitz, P. 1968. Applications of linear programming to numerical analysis. *SIAM-REV* **10**, 121–159.

Rabinowitz, P. (ed.). 1970. *Numerical Methods for Nonlinear Algebraic Equations*. London: Gordon and Breach.

Raimi, R. A. 1969. On the distribution of first significant figures. *AMM* **76**, 342–347.

Rall, L. B. 1965. *Error in Digital Computation*. New York: Wiley.

Ralston, A., and C. L. Meek (eds.). 1976. *Encyclopedia of Computer Science*. New York: Petrocelli/Charter.

Ralston, A., and P. Rabinowitz. 1978. *A First Course in Numerical Analysis*. New York: McGraw-Hill.

Rand, R. 1984. *Computer Algebra in Applied Mathematics: An Introduction to* . Boston: Pitman.

Redish, K. A. 1974. On Laguerre's method. *Int. J. Math. Educ. Sci. Technol.* **5**, 91–102.

Reid, J. K. (ed.). 1971. *Large Sparse Sets of Linear Equations*. New York: Academic Press.

Renka, R. J. 1993. Algorithm 716: TSPACK–Tension spline curve-fitting package. *ACM-TOMS* **19**, 81–94.

Rheinboldt, W. C. 1974. *Methods for Solving Systems of Nonlinear Equations*. CBMS Series in Applied Mathematics 14. Philadelphia: SIAM.

Rheinboldt, W. C. 1980. Solution fields of nonlinear equations and continuation methods. *SIAM-NA* **17**, 221–237.

Rheinboldt, W. C. 1986. *Numerical Analysis of Parameterized Nonlinear Equations*. New York: Wiley.

Rice, J. R. 1966. Experiments on Gram-Schmidt orthogonalization. *MOC* **20**, 325–328.

Rice, J. R. 1981. *Matrix Computations and Mathematical Software*. New York: McGraw-Hill.

Rice, J. R. 1992. *Numerical Methods, Software, and Analysis*. 2nd ed. New York: Academic Press.

Rice, J. R., and R. F. Boisvert. 1985. *Elliptic Problem Solving Using* . New York: Springer-Verlag.

Rice, J. R., and J. S. White. 1964. Norms for smoothing and estimation. *SIAM-REV* **6**, 243–256.

Richtmeyer, R. D., and K. W. Morton. 1967. *Difference Methods for Initial Value Problems*. New York: Wiley.

Rivlin, T. J. 1990. *The Chebyshev Polynomials*. 2nd ed. New York: Wiley.

Roache, P. 1972. *Computational Fluid Dynamics*. Albuquerque, NM: Hermosa.

Roberts, S., and J. Shipman. 1972. *Two-Point Boundary Value Problems: Shooting Methods*. New York: Elsevier.

Rockafellar, R. T. 1970. *Convex Analysis*. Princeton, NJ: Princeton University Press.

Rose, D. J. 1975. A simple proof for partial pivoting. *AMM* **82**, 919–921.

Rose, D. J., and R. A. Willoughby (eds.). 1972. *Sparse Matrices and Their Applications*. New York: Plenum.

Rosenfeld, A., and A. Kak. 1982. *Digital Picture Processing*. New York: Academic Press.

Ross, S. 1983. *Stochastic Processes*. New York: Wiley.

Rousseau, C. 1995. The phi number system revisited. *Math. Mag.* **68**, 283–284.

Roy, M. R. 1985. *A History of Computing Technology*. Englewood Cliffs, NJ: Prentice-Hall.

Royden, H. L. 1968. *Real Analysis*. 2nd ed. New York: Macmillan.

Rozema, E. R. 1987. Romberg integration by Taylor series. *AMM* **94**, 284–288.

Rubinstein, R. 1981. *Simulation and the Monte Carlo Method*. New York: Wiley.

Ryder, B. G. 1974. The PFORT verifier. *Software Practice and Experience* **4**, 359–378.

Saaty, T. L. 1981. *Modern Nonlinear Equations*. New York: Dover.

Salamin, E. 1976. Computation of π using arithmetic-geometric mean. *MOC* **30**, 565–570.

Sander, L. M. 1987. Fractal growth. *SA*, January **256**, 94–100.

Sard, A. 1963. *Linear Approximation*. Mathematical Surveys, No. 9. Providence, RI: AMS.

Scales, L. E. 1985. *Introduction to Non-Linear Optimization*, New York: Macmillan.

Schechter, M. 1984. Summation of divergent series by computer. *AMM* **91**, 629–632.

Scheid, F. 1988. *Numerical Analysis*. New York: McGraw-Hill.

Schendel, U. 1984. *Introduction to Numerical Methods for Parallel Computers*. New York: Wiley.

Schiesser, W. E. 1994. *Computational Mathematics in Engineering and Applied Science*. Boca Raton, FL: CRC Press.

Schnabel, R. B., and P. D. Frank. 1984. Tensor methods for nonlinear equations. *SIAM-NA* **21**, 815–843.

Schnabel, R. B., J. E. Koontz, and B. E. Weiss. 1982. A modular system of algorithms for unconstrained minimization. Computer Science Department Report CU-CS-240-82. Boulder, CO: University of Colorado at Boulder.

Schoenberg, I. J: 1946. Contributions to the problem of approximation of equidistant data by analytic functions. *Quarterly Applied Mathematics* **4**, 45–99 and 112–133.

Schoenberg, I. J. 1967. On spline functions. In *Inequalities* (O. Shisha, ed.), 255–291. New York: Academic Press.

Schoenberg, I. J. 1982. *Mathematical Time Exposures*. Washington, DC: MAA.

Schrage, L. 1979. A more portable Fortran random number generator. *ACM-TOMS* **5**, 132–138.

Schultz, M. H. 1973. *Spline Analysis*. Englewood Cliffs, NJ: Prentice-Hall.

Schumaker, L. L. 1976. Fitting surfaces to scattered data. In *Approximation Theory II* (G. G. Lorentz, C. K. Chui, and L. L. Schumaker, eds.), 203–268. New York: Academic Press.

Schumaker, L. L. 1981. *Spline Functions*. New York: Wiley-Interscience.

Schweikert, D. G. 1966. An interpolation curve using splines in tension. *JMP* **45**, 312–317.

Scott, N. R. 1985. *Computer Number Systems and Arithmetic*. Englewood Cliffs, NJ: Prentice-Hall.

Shampine, L. F. 1994. *Numerical Solution of Ordinary Differential Equations*. New York: Chapman and Hall.

Shampine, L. F., and R. C. Allen. 1973. *Numerical Computing: An Introduction*. Philadelphia: Saunders.

Shampine, L. F., and C. Baca. 1984. Error estimators for stiff differential equations. *JCAM* **2**, 197–208.

Shampine, L. F., and P. Bogacki. 1989. The effect of changing the stepsize in linear multistep codes. *SIAM-SSC* **10**, 1010–1023.

Shampine, L. F., and C. W. Gear. 1979. A user's view of solving stiff ordinary differential equations. *SIAM-REV* **21**, 1–17.

Shampine, L. F., and M. K. Gordon. 1975. *Computer Solution of Ordinary Differential Equations: The Initial Value Problem*. San Francisco: Freeman.

Shampine, L. F., H. A. Watts, and S. M. Davenport. 1976. Solving nonstiff ordinary differential equations—The state of the art. *SIAM-REV* **18**, 376–411.

Shepard, D. 1968. A two-dimensional interpolation function for irregularly spaced data. *Proceedings 23rd National Conference ACM*, 517–524.

Shikin, E. V. 1995. *Handbook and Atlas of Curves*. Boca Raton, FL: CRC Press.

Skeel, R. D. 1979. Equivalent forms of multistep formulas. *MOC* **33**, 1229–1250.

Skeel, R. D. 1981. Effect of equilibration on residual size for partial pivoting. *SIAM-NA* **18**, 449–454.

Sloan, I. H., and S. Joe. 1994. *Lattice Methods for Multiple Integration*. New York: Oxford University Press.

Smale, S. 1981. The fundamental theorem of algebra and complexity theory. *AMS-B* **4**, 1–36.

Smale, S. 1986. Algorithms for solving equations. In *Proceedings International Congress of Mathematicians* (A. M. Gleason, ed.), 172–195, Providence, RI: AMS.

Smith, B. T., J. M. Boyle, B. S. Garbow, Y. Ikebe, V. C. Klema, and C. B. Moler. 1976. *Matrix Eigensystem Routines— Guide*. 2nd ed.: *LNCS* **6**. New York: Springer-Verlag.

Smith, G. D. 1965. *Numerical Solution of Partial Differential Equations*. New York: Oxford University Press.

Smith, K. T. 1971. *Primer of Modern Analysis*. New York: Springer-Verlag.

Sobolev, S. L. 1992. *Cubature Formulas and Modern Analysis*. Philadelphia: Gordon and Breach.

Sorenson, DC 1985. Analysis of pairwise pivoting in Gaussian elimination. *IEEE-TC* **34**, 274–278.

Steffensen, J. F. 1950. *Interpolation*. 2nd ed. New York: Chelsea.

Steinberg, D. I. 1975. *Computational Matrix Algebra*. New York: McGraw-Hill.

Sternbenz, P. H. 1974. *Floating-Point Computations*. Englewood Cliffs, NJ: Prentice-Hall.

Stetter, H. J. 1973. *Analysis of Discretization Methods for Ordinary Differential Equations*. New York: Springer-Verlag.

Stewart, G. W. 1973. *Introduction to Matrix Computations*. New York: Academic Press.

Stewart, G. W. 1985. A note on complex division. *ACM-TOMS* **11**, 238–341.

Stoer, J., and R. Bulirsch. 1980. *Introduction to Numerical Analysis*. New York: Springer-Verlag.

Strang, G., and G. Fix. 1973. *An Analysis of the Finite Element Method*. Englewood Cliffs, NJ: Prentice-Hall.

Street, R. L. 1973. *The Analysis and Solution of Partial Differential Equations*. Pacific Grove, CA: Brooks/Cole.

Stroud, A. H. 1965. Error estimates for Romberg quadrature. *SIAM-NA* **2**, 480–488.

Stroud, A. H. 1971. *Approximate Calculation of Multiple Integrals*. Englewood Cliffs, NJ: Prentice-Hall.

Stroud, A. H. 1974. *Numerical Quadrature and Solution of Ordinary Differential Equations*. New York: Springer-Verlag.

Stroud, A. H., and D. Secrest. 1966. *Gaussian Quadrature Formulas*. Englewood Cliffs, NJ: Prentice-Hall.

Subbotin, Y. N. 1967. On piecewise-polynomial approximation. *Math. Zametcki* **1**, 63–70. (Transl.: 1967. *Mathematical Notes* **1**, 41–46.)

Swarztrauber, P. N. 1975. Efficient subprograms for the solution of elliptic partial differential equations. Report TN/LA-109. Boulder, CO: National Center for Atmospheric Research.

Swarztrauber, P. N. 1982. Vectorizing the FFT's parallel computations. In *Parallel Computations* (G. Rodrigue, ed.). New York: Academic Press.

Swarztrauber, P. N. 1984. Fast Poisson solvers. In *Studies in Numerical Analysis* (G. H. Golub, ed.), 319–370. Washington, DC: MAA.

Taussky, O. 1949. A remark concerning the characteristic roots of finite segments of the Hilbert matrix. *Oxford Quarterly Journal Mathematics* **20**, 82–83.

Taussky, O. 1988. How I became a torchbearer for matrix theory. *AMM* **95**, 801–812.

Taylor, J. R. 1982. *An Introduction to Error Analysis*. New York: University Science Books.

Tewarson, R. P. 1973. *Sparse Matrices*. New York: Academic Press.

Thomas, B. 1986. The Runge-Kutta methods. *Byte*, April, 191–210.

Todd, J. 1961. Computational problems concerning the Hilbert matrix. *JR-NBS* **65**, 19–22.

Todd, M. J. 1982. An introduction to piecewise linear homotopy algorithms for solving systems of equations. In *Topics in Numerical Analysis* (P. R. Turner, ed.) *LNM* **965**, 147–202. New York: Springer-Verlag.

Torczon, V. 1997. On the convergence of pattern search methods, *SIAM-JO* **7**, 1–25.

Törn, A., and A. Zilinska. 1989. *Global Optimization. LNCS* **350**. New York: Springer-Verlag.

Traub, J. F. 1964. *Iterative Methods for the Solution of Equations*. Englewood Cliffs, NJ: Prentice-Hall.

Traub, J. F. 1967. The calculation of zeros of polynomials and analytic functions. In *Mathematical Aspects of Computer Science. Proceedings Symposium Applied Mathematics* **19**, 138–152. Providence, RI: AMS.

Trefethen, L. N. 1992. The definition of numerical analysis. Report TR 92–1304. Ithaca, NY: Cornell University.

Trefethen, L. N., and R. S. Schreiber. 1990. Average-case stability of Gaussian elimination. *SIAM-MAA* **11**, 335–360.

Trustrum, K. 1971. *Linear Programming*. London: Routledge and Kegan Paul.

Tseng, P. 1998. Fortified-descent simplicial search method: a general approach, preprint.

Turner, P. R. 1982. The distribution of leading significant digits. *IMA-JNA* **2**, 407–412.

van der Corput, J. G. 1946. Sur l'approximation de Laguerre des racines d'une équation qui a toutes ses racines réelles. *Nederl. Acad. Wetensch. Proc.* **49**, 922–929.

Vandergraft, J. S. 1978. *Introduction to Numerical Computations*. New York: Academic Press.

van Huffel, S., and J. Vandervalle. 1991. *The Total Least Squares Problem: Computational Aspects and Analysis*. Philadelphia: SIAM.

van Loan, C. F. 1992. *Computational Frameworks for the Fast Fourier Transform*. Philadelphia: SIAM.

Varga, R. S. 1962. *Matrix Iterative Analysis*. Englewood Cliffs, NJ: Prentice-Hall. (2000. *Matrix Iterative Analysis: Second Revised and Expanded Edition*. New York: Spring-Verlag.)

Vemuri, V., and W. J. Karplus. 1981. *Digital Computer Treatment of Partial Differential Equations*. Englewood Cliffs, NJ: Prentice-Hall.

Verner, J. H. 1978. Explicit Runge-Kutta methods with estimates of the local truncation error. *SIAM-NA* **15**, 772–790.

Vichnevetsky, R. 1981. *Computer Methods for Partial Differential Equations*. Vol. 1: *Elliptic Equations and the Finite Element Method*, 1982. Vol. 2: *Initial Value Problems*. Englewood Cliffs, NJ: Prentice-Hall.

Von Petersdorff, T. 1993. A short proof for Romberg integration. *AMM* **100**, 783–785.

Von Rosenberg, D. U. 1969. *Methods for the Numerical Solution of Partial Differential Equations*. New York: American Elsevier.

Von Seggern, D. 1994. *Practical Handbook of Curve Design and Generation*. Boca Raton, FL: CRC Press.

Wachspress, E. L. 1966. *Iterative Solution of Elliptic Systems*. Englewood Cliffs, NJ: Prentice-Hall.

Wacker, H. G. (ed.). 1978. *Continuation Methods*. New York: Academic Press.

Wait, R., and A. R. Mitchell. 1986. *Finite Element Analysis and Applications*. New York: Wiley.

Walker, J. S. 1992. *Fast Fourier Transforms*. Boca Raton, FL: CRC Press.

Wall, H. S. 1948. *Analytic Theory of Continued Fractions*. Princeton: van Nostrand.

Walsh, G. R. 1975. *Methods of Optimization*, New York: Wiley.

Waser, S., and M. J. Flynn. 1982. *Introduction to Arithmetic for Digital Systems Designers*. New York: Holt, Reinhart & Winston.

Wasserstrom, E. 1973. Numerical solutions by the continuation method. *SIAM-REV* **15**, 89–119.

Watkins, D. S. 1982. Understanding the *QR* algorithm. *SIAM-REV* **24**, 427–440.

Watson, L. T. 1979. A globally convergent algorithm for computing fixed points of C^2 maps. *Applications in Mathematical Computing* **5**, 297–311.

Watson, L. T. 1986. Numerical linear algebra aspects of globally convergent homotopy methods. *SIAM-REV* **28**, 529–545.

Watson, L. T., S. C. Billups, and A. P. Morgan. 1987. HOMPACK: A suite of codes for globally convergent homotopy algorithms. *ACM-TOMS* **13**, 281–310.

Weaver, H. J. 1983. *Applications of Discrete and Continuous Fourier Analysis*. New York: Wiley.

Wedin, P. A. 1972a. Perturbation bounds in connection with the singular value decomposition. *BIT* **12**, 99–111.

Wedin, P. A. 1972b. Perturbation theory for pseudoinverses. *BIT* **13**, 217–232.

Werner, W. 1984. Polynomial interpolation: Lagrange versus Newton. *MOC* **43**, 205–217.

Wesseling, P. 1992. *An Introduction to Multigrid Methods*. New York: Wiley.

Westfall, R. S. 1980. *Never at Rest: A Biography of Isaac Newton*. New York: Cambridge University Press.

Whitehead, G. W. 1966. *Homotopy Theory*. Cambridge, MA: MIT Press.

Whittaker, E., and G. Robinson. 1924. *The Calculus of Observations*. 4th ed. London: Blackie and Son. (Reprinted New York: Dover, 1967.)

Wilkinson, J. H. 1961. Error analysis of direct methods of matrix inversion. *ACM-J* **8**, 281–330.

Wilkinson, J. H. 1963. *Rounding Errors in Algebraic Processes*. Englewood Cliffs, NJ: Prentice-Hall.

Wilkinson, J. H. 1965. *The Algebraic Eigenvalue Problem*. New York: Oxford University Press.

Wilkinson, J. H. 1967. Two algorithms based on successive linear interpolation. Technical Computer Science Department Report STAN-CS-67-60. Stanford, CA: Stanford University.

Wilkinson, J. H. 1984. The perfidious polynomial. In *Studies in Numerical Analysis* (G. H. Golub, ed.), 1–28. Washington, DC: MAA.

Wilkinson, J. M., and C. Rheinsch (eds.). 1971. *Handbook for Automatic Computation II: Linear Algebra*. New York: Springer-Verlag.

Willé, D. R. 1989. The numerical solution of delay-differential equations. Department of Mathematics, Ph.D. thesis. Manchester, England: University of Manchester.

Willé, D. R. 1994a. New stepsize estimators for linear multistep methods. Department of Mathematics, Numerical Analysis Report No. 247. Manchester, England: University of Manchester.

Willé, D. R. 1994b. Experiments in stepsize control for Adams linear multistep methods. Department of Mathematics, Numerical Analysis Report No. 253. Manchester, England: University of Manchester.

Willé, D. R., and C. T. H. Baker. 1992. DELSOL: A numerical code for the solution of systems of delay-differential equations. *ANM* **9**, 223–234.

Willoughby, R. A. (ed.). 1974. *Stiff Differential Systems*. New York: Plenum.

Wimp, J. 1984. *Computation with Recurrence Relations*. Boston: Pitman.

Wouk, A. (ed.). 1986. *New Computing Environments: Parallel, Vector, and Systolic*. Philadelphia: SIAM.

Wright, M. 1991. Interior methods for constrained optimization. *AN* 341–407.

Wright, S. J. 1993. A collection of problems for which Gaussian elimination with partial pivoting is unstable. *SIAM-SSC* **14**, 231–238.

Young, D. M. 1950. Iterative methods for solving partial difference equations of elliptic type. Ph.D. thesis, Cambridge, MA: Harvard University.

Young, D. M. 1971. *Iterative Solution of Large Linear Systems*. New York: Academic Press.

Young, D. M., and R. T. Gregory. 1972. *A Survey of Numerical Mathematics*. Vols. 1 and 2. Reading, MA: Addison-Wesley. (Reprinted New York: Dover, 1988.)

Young, D. M., and K. C. Jea. 1980. Generalized conjugate acceleration of nonsymmetrizable iterative methods. *LAA* **34**, 159–194.

Young, R. M. 1986. A Rayleigh popular problem. *AMM* **93**, 660.

Zelkowitz, M. V., A. C. Shaw, and J. D. Gannon. 1979. *Principles of Software Engineering and Design*. Englewood Cliffs, NJ: Prentice-Hall.

Zienkiewicz, O. C., and K. Morgan. 1983. *Finite Elements and Approximation*. New York: Wiley.

Zwillinger, D. 1988. *Handbook of Differential Equations*. New York: Academic Press.

索　引

索引中的页码为英文原书页码，与书中页边标注的页码一致.

B

C

E

H

J

K

L

M

N

O

Q

R

S

T

U